ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία

ΣΥΓΚΡΙΣΗ ΛΟΓΙΣΜΙΚΩΝ ΑΝΟΙΚΤΟΥ ΚΩΔΙΚΑ ΕΞΟΡΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
Η ΜΕΘΟΔΟΣ ΤΗΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Του φοιτητή
Ζαρμακούπη Θεοφάνη
Αρ.Μητρώου: 03/2302

Επιβλέπων Καθηγητής
Καραμητόπουλος Λεωνίδας

Θεσσαλονίκη 2011

ΠΕΡΙΛΗΨΗ

Στην σημερινή εποχή η απόκτηση της πληροφορίας και η εξόρυξη γνώσης από αυτή παίζει πολύ σημαντικό ρόλο. Ο τομέας της εξόρυξης δεδομένων αποκτά όλο και μεγαλύτερη αξία μιας και η εξέλιξη της τεχνολογίας παρέχει καινούριες δυνατότητες. Στην παρούσα εργασία εξηγείται τι είναι το Data Mining, ποιες είναι οι βασικές ενέργειες του και αναφέρονται κάποιες μέθοδοι τους. Αναλύεται η συσταδοποίηση και ένας από τους σημαντικότερους αλγορίθμους της, ο k-means. Στην συνέχεια γίνεται χρήση μιας μελέτης περίπτωσης που αφορά μια τράπεζα και τους πελάτες της και εφαρμόζεται πάνω σ αυτή ο k-means σε τρία δωρεάν προγράμματα: το Weka, το Orange και το Rapidminer. Αναφέρονται επίσης όλα τα προγράμματα είτε είναι δωρεάν, είτε όχι.

Η πτυχιακή εργασία αυτή εστιάζεται κυρίως στην ανάλυση των αποτελεσμάτων και γίνεται σύγκριση των προγραμμάτων και των λειτουργιών τους, έτσι ώστε ο αναγνώστης να σχηματίσει άποψη και να επιλέξει όποιο εργαλείο προτιμά. Είναι πολύ σημαντικό το ότι χρησιμοποιούνται τα ίδια δεδομένα και στα τρία εργαλεία και μπορεί να γίνει καλύτερη σύγκριση αποτελεσμάτων.

ABSTRACT

Nowadays the acquisition of information and knowledge mining from that has a very important role. The sector of data mining gets increasing value and the development of technology provides new abilities. In the present paper, it is explained what is Data Mining, which are the main activities and some of the methods of them are referred. Clustering is analyzed and one of the most important algorithms of it, k-means. Then a case study that is about a bank and its customers is used and is applied on this case study the k-means algorithm on three free programs: Weka, Orange and Rapidminer. All programs are also referred in spite whether are free or not.

This paper focuses mainly on the analysis of results and the comparison of three programs and their functions, so the reader can make an opinion and choose which program that he favors. It is very important that the same data are used to all three programs and can be done a better comparison.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή κ. Λεωνίδα Καραμητόπουλο για την πολύ καλή συνεργασία που είχαμε αλλά και τις συμβουλές του.

Επίσης θα ήθελα να ευχαριστήσω την οικογένειά μου και κυρίως τους γονείς μου για την αμέριστη συμπαράστασή τους όλα αυτά τα χρόνια.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1
Εισαγωγή

1.1 Πρόλογος

Οι συνεχείς εξελίξεις στον τομέα των ηλεκτρονικών υπολογιστών, στο επίπεδο της ανάπτυξης λογισμικού (software) και hardware δίνουν συνεχώς νέες δυνατότητες στα ηλεκτρονικά συστήματα για επιπλέον χρήση τους σε τομείς που ως τώρα οι πεπερασμένοι πόροι τους δυσχέραιναν την ανάπτυξη νέων εργαλείων. Στην σημερινή κοινωνία της πληροφορίας, υπάρχουν συνεχώς αυξανόμενοι όγκοι δεδομένων (εικόνες, βίντεο, ήχος κτλ) που με περισσότερο εξελιγμένο hardware μπορούν πλέον πιο εύκολα και να αποθηκευτούν λόγω μεγαλύτερης χωρητικότητας, αλλά και να επεξεργαστούν (ταχύτεροι επεξεργαστές, μνήμες κτλ). Με τις συνεχείς καταγραφές όλων των τύπων δεδομένων, δημιουργούνται πολύ μεγάλες βάσεις δεδομένων και έτσι πρέπει να βρεθεί ένας τρόπος να πάρουμε την χρήσιμη πληροφορία από αυτά τα δεδομένα αποτελεσματικά και αποδοτικά.

Τα τελευταία χρόνια, έχουν αναπτυχθεί όλο και πιο αποδοτικές τεχνικές ανάλυσης δεδομένων που βοηθούν προς την κατεύθυνση αυτή, αλλά και πολλά εργαλεία που τις χρησιμοποιούνε. Ο τομέας της Πληροφορικής που ασχολείται με το θέμα αυτό είναι το Data mining (Εξόρυξη Πληροφορίας) και έχει ευρεία εφαρμογή. Στο παρόν σύγγραμμα θα επικεντρωθούμε σε εργαλεία που μας βοηθούν να πάρουμε την πληροφορία-γνώση από βάσεις δεδομένων, εργαλεία δηλαδή που έχουν ενσωματωμένες τεχνικές Data Mining. Τα εργαλεία που επιλέχτηκαν είναι open source και θα γίνει σύγκριση τους με διάφορα κριτήρια όπως αποτελεσματικότητας, αποδοτικότητας και ακρίβειας.

1.2 Σκοπός και ερευνητικές ερωτήσεις

Η παρούσα εργασία αποσκοπεί στον έλεγχο της αποτελεσματικότητας και αποδοτικότητας τριών διαφορετικών open source εργαλείων για την εξόρυξη της πληροφορίας (data mining) και πιο συγκεκριμένα για την συσταδοποίηση (clustering) και την εφαρμογή του πολύ δημοφιλούς αλγόριθμου k-means. Τα εργαλεία είναι το Weka 3.6.4, το Orange (version 2.0b), και το Rapidminer 5.1.006.

Η σύγκριση θα γίνει σε όλα τα προγράμματα με το ίδιο σύνολο δεδομένων, το οποίο αφορά σε μία επιχειρηματική περίπτωση.

Θα μελετηθεί ο τρόπος λειτουργίας του κάθε εργαλείου και θα αναλυθούν τα αποτελέσματα που θα βγούνε για το ίδιο σύνολο δεδομένων. Η εργασία αυτή θα απαντήσει σε ερωτήματα όπως:

- ✓ Ποιο εργαλείο από τα τρία είναι πιο εύχρηστο;
- ✓ Ποιο εργαλείο είναι πιο φιλικό προς τον χρήστη;
- ✓ Ποιο εργαλείο επιστρέφει πληρέστερα και σαφέστερα αποτελέσματα;
- ✓ Ποιο εργαλείο παρέχει αποτελέσματα με την μεγαλύτερη ακρίβεια;
- ✓ Ποιο εργαλείο είναι πιο έγκυρο;
- ✓ Ποιο εργαλείο έχει λιγότερα προβλήματα;
- ✓ Ποιο εργαλείο παρέχει περισσότερες επιλογές στις παραμέτρους των μεθόδων;
- ✓ Ποιο εργαλείο έχει καλύτερη απεικόνιση αποτελεσμάτων;
- ✓ Ποιο εργαλείο έχει τους περισσότερους αλγόριθμους για συσταδοποίηση;

1.3 Διάρθρωση εργασίας

Στο δεύτερο κεφάλαιο αναπτύσσονται οι βασικές έννοιες της Εξόρυξης Πληροφορίας, αναφέρονται σχετικά λογισμικά πακέτα, όπως επίσης και περιγράφονται τα 3 λογισμικά που θα συγκριθούν. Στο τρίτο κεφάλαιο αναλύεται θεωρητικά η τεχνική της συσταδοποίησης και οι αντίστοιχοι αλγόριθμοι. Επίσης, γίνεται εκτενής περιγραφή του αλγορίθμου k-means ο οποίος και θα εφαρμοστεί στην μελέτη περίπτωσης. Στο τέταρτο κεφάλαιο θα παρουσιαστεί η μελέτη περίπτωσης. Συγκεκριμένα, θα περιγραφεί ο σκοπός της μελέτης, τα χαρακτηριστικά (μεταβλητές) εκείνα για τα οποία υπάρχουν καταγεγραμμένες παρατηρήσεις και θα εξηγηθεί για ποιον λόγο η τεχνική της συσταδοποίησης μπορεί να απαντήσει στον σκοπό της μελέτης. Στο πέμπτο κεφάλαιο θα παρουσιαστεί η εφαρμογή του αλγορίθμου k-means σε κάθε λογισμικό και θα εξηγηθεί αναλυτικά η διαδικασία που ακολουθείται. Επίσης, θα δοθούν κατάλληλες ερμηνείες των αποτελεσμάτων για κάθε ένα από τα προγράμματα. Στο έκτο και τελευταίο κεφάλαιο θα πραγματοποιηθεί σύγκριση των λογισμικών με βάση τα γενικά χαρακτηριστικά τους και την διαδικασία εφαρμογής της συσταδοποίησης.

ΚΕΦΑΛΑΙΟ 2
Εξόρυξη πληροφορίας

2.1 Εισαγωγή

Καθημερινά ο άνθρωπος δέχεται σωρεία δεδομένων τα οποία συλλέγει και μέσω της σκέψης και της κατανόησής τους ξεχωρίζει την πληροφορία που εμπεριέχεται σε αυτά. Ας ξεχωρίσουμε λοιπόν πρώτα απ' όλα την διαφορά μεταξύ δεδομένου και πληροφορίας. «Τα **δεδομένα** (**data**), είναι γεγονότα, μηνύματα, κωδικοποιημένα ή όχι και αποτελούν ακατέργαστο πληροφοριακό υλικό»[1].Είναι δηλαδή η εισροή στην διαδικασία της επεξεργασίας με αποτέλεσμά της την πληροφορία. **«Πληροφορία** είναι το κοινό νόημα που αποδίδεται σε ένα απλό ή σύνθετο σύμβολο από δύο ή περισσότερα υποκείμενα[2]».

Σήμερα με την συνεχή ανάπτυξη hardware είναι δυνατή η αποθήκευση αλλά και η επεξεργασία πολύ μεγάλων όγκων δεδομένων που μέχρι πριν κάποια χρόνια ήταν πολύ ασύμφορο να γίνει λόγω κόστους των πόρων που απαιτούνταν. Σήμερα αναπτύσσεται ένας σχετικά καινούριος τομέας της Πληροφορικής που λέγεται Εξόρυξη Πληροφορίας (Data Mining). «Η **Εξόρυξη Πληροφορίας (Data Mining)** είναι ένας σχετικά νέος και διεπιστημονικός τομέας της επιστήμης των υπολογιστών, ο οποίος αποτελεί την διαδικασία αναγνώρισης μοτίβων (patterns) από μεγάλα σύνολα δεδομένων (dataset) συνδυάζοντας μεθόδους από την στατιστική, την τεχνητή νοημοσύνη και την διαχείριση βάσεων δεδομένων» [3]. «Η σύνθετη διαδικασία εξαγωγής συγκεκριμένης, προηγουμένως άγνωστης και δυνητικά ωφέλιμης, γνώσης από δεδομένα»[4]. Εναλλακτικά, συναντάται και ως «η επιστήμη της εξόρυξης χρήσιμης πληροφορίας από βάσεις δεδομένων μεγάλου μεγέθους» [5]. «Αναφορικά με τη διαχείριση επιχειρηματικών πόρων (ERP), το **Data Mining** θεωρείται ως η στατιστική και λογική ανάλυση εκτεταμένων συνόλων από δεδομένα συναλλαγών και εργασιών για τον εντοπισμό επαναλαμβανόμενων μοτίβων ή τάσεων που μπορούν να βοηθήσουν στη λήψη αποφάσεων» [6]. Για παράδειγμα, έχοντας μια βάση δεδομένων και με χρήση της κατάλληλης μεθόδου μπορούν να συσχετιστούν κάποια δεδομένα μεταξύ τους, ακόμα και να ‘’προβλεφθούν’’ με χρήση πιθανοτήτων κάποια αποτελέσματα. Είναι εύκολα

αντιληπτό ότι το Data Mining αποτελεί μια από τις αναδυόμενες τεχνολογίες που φέρνει μεγάλες εξελίξεις στον χώρο της ανάλυσης δεδομένων.

2.2 Ενέργειες & Μέθοδοι

Θα αναφέρουμε και θα εξηγήσουμε κάποιες βασικές ενέργειες (tasks) Data mining καθώς και κάποιες μεθόδους τους.

Κατηγοριοποίηση (classification) και πρόβλεψη: Χρησιμοποιείται στο να κατασκευάζουμε μοντέλα που περιγράφουν και διαχωρίζουν κάποια αντικείμενα σε μία σειρά κατηγοριών, οι οποίες είναι γνωστές εκ των προτέρων. Ο στόχος είναι να περιγραφούν τα δεδομένα ή να γίνει μια μελλοντική κατηγοριοποίηση (πρόβλεψη), όπως για παράδειγμα η κατηγοριοποίηση των πελατών μίας εταιρείας σε πελάτες «υψηλού» ή «χαμηλού» κινδύνου με βάση τα διάφορα χαρακτηριστικά τους. Οι τρεις από τις περισσότερο δημοφιλείς μεθόδους είναι τα Δέντρα Αποφάσεων (decision trees), οι Κανόνες Κατηγοριοποίησης (classification rules) και τα Νευρωνικά Δίκτυα (neural networks). Σε ορισμένες μεθόδους είναι δυνατή και η πρόβλεψη των αριθμητικών τιμών κάποιων χαρακτηριστικών.

Συσταδοποίηση(clustering): Χρησιμοποιείται στο να κατασκευάζουμε μοντέλα που περιγράφουν και διαχωρίζουν κάποια αντικείμενα σε μία σειρά κατηγοριών (συστάδων), οι οποίες είναι άγνωστες εκ των προτέρων. Παράδειγμα συσταδοποίσης αποτελεί ο διαχωρισμός των καταναλωτών σε μία σειρά ομάδων (συστάδων) με βάση τα δημογραφικά στοιχεία τους. Η συσταδοποίηση έχει ως στόχο να δημιουργεί ομάδες με αυξημένη ομοιότητα μεταξύ αντικειμένων της ίδιας συστάδας και μειωμένη ομοιότητα μεταξύ αντικειμένων διαφορετικών συστάδων.

Κανόνες συσχέτισης(Association rules): Με αυτούς βρίσκουμε ενδιαφέροντες σχέσεις και συσχετίσεις μέσα σε μεγάλα σύνολα δεδομένων. Παράδειγμα ενός τέοιου κανόνα είναι το εξής: τα άτομα εκείνα που αγοράζουν διάφορα αξεσουάρ για το αυτοκίνητό τους, έχουν την τάση να καθυστερούν την πληρωμή της ασφάλειας του αυτοκινήτου τους.

Ανάλυση ακραίων τιμών(Outlier analysis): Outlier είναι ένα αντικείμενο που δεν ταιριάζει με την γενική συμπεριφορά των δεδομένων. Μπορεί να θεωρηθεί ως θόρυβος ή εξαίρεση αλλά είναι πολύ χρήσιμο στην αναγνώριση απάτης και σε ανάλυση σπάνιων φαινομένων.

Ανάλυση τάσης και εξέλιξης (Trend and evolution analysis): Χρησιμεύει στην αναγνώριση μοτίβων ακολουθίας, ανάλυση περιοδικότητας και αναλύσεις σε σχέση με την ομοιότητα.

Υπάρχουν κι άλλες μέθοδοι όπως το Text mining, Graph mining και άλλες στατιστικές τεχνικές που χρησιμοποιούνται με πιο διαδομένες αυτές που περιγράφονται παραπάνω.[19]

2.3 Λογισμικά πακέτα

Υπάρχει μια πληθώρα λογισμικών για την εξόρυξη της πληροφορίας και χωρίζονται σε εμπορικά και σε ανοικτού κώδικα. Τα εμπορικά δεν είναι δωρεάν σε αντίθεση με τα λογισμικά ανοικτού κώδικα. Παρακάτω παρατίθενται κάποια εμπορικά και κάποια δωρεάν προγράμματα που είναι πιο δημοφιλή.

Α) **Εμπορικά**

- **SAS** (Statistical Analysis System*)* **Enterprise Miner** – λογισμικό που παρέχεται από το SAS Institute.
- **SPSS Modeler** – λογισμικό που παρέχεται από την IBM SPSS. Σύμφωνα με το Rexer's Annual Data Miner Survey το 2010, το πρόγραμμα μαζί με το STATISTICA Data Miner και το R) παρείχε την υψηλότερη ικανοποίηση στους χρήστες και το 2010 και το 2009.[20]
- **STATISTICA Data Miner** – λογισμικό από την StatSoft. Όπως αναφέρεται και στο προηγούμενο πρόγραμμα παρείχε την υψηλότερη ικανοποίηση στου χρήστες το 2009 και 2010. Επιπλέον το 2010 βαθμολογήθηκε ως το πιο συχνά επιλεχθέν βασικό εργαλείο data mining (18%).[20]
- **DB2 Intelligent Miner** – λογισμικό από την IBM. Επιτρέπει και σε SQL εφαρμογές να διαχειριστούν μεθόδους εξόρυξης πληροφορίας.
- **Microsoft Analysis Services** – λογισμικό από την Microsoft. Είναι κομμάτι του Microsoft SQL Server που είναι ένα σύστημα διαχείρισης βάσεων. Παρέχει και κάποιες δυνατότητες εξόρυξης πληροφορίας.

Β) **Ανοικτού Κώδικα**

- **ELKI** – ένα πανεπιστημιακό εργαλείο με ανάλυση συστάδας(cluster analysis) και εύρεση ακραίων τιμών(outlier detection) γραμμένο σε Java.
- **KNIME** – The Konstanz Information Miner, ένα φιλικό για τον χρήστη και περιεκτικό περιβάλλον για την ανάλυση δεδομένων.
- **Orange** – ένα machine learning περιβάλλον που αποτελείται από components(κομμάτια).
- **R** – Μια γλώσσα προγραμματισμού και περιβάλλον για υπολογισμό στατιστικών, εξόρυξη πληροφορίας και γραφικά. Είναι κομμάτι του GNU project. Το 2010 ήταν το πιο πολύ χρησιμοποιημένο από χρήστες για εξόρυξη πληροφορίας από οποιοδήποτε άλλο.[20]
- **RapidMiner** – Ένα machine learning περιβάλλον για πειράματα εξόρυξης πληροφορίας.
- **Weka** – Ένα machine learning περιβάλλον γραμμένο σε Java.
- **Carrot2** – Ένα framework για συσταδοποίηση σε αποτελέσματα κειμένων και ευρετηρίων.
- **GATE** – Ένα εργαλείο για επεξεργασία φυσικής γλώσσας.
- **JHepWork** – Ένα εργαλείο σε Java για ανάλυση δεδομένων που έγινε από την ANL.
- **NLTK** – Ένα σύνολο βιβλιοθηκών και προγραμμάτων για επεξεργασία των στατιστικών φυσικής γλώσσας σε Python.
- **Tanagra** – Ένα εργαλείο που εμπεριέχει ανάλυση δεδομένων αλγορίθμους για στατιστικά.
- **Text Engineering Software Laboratory** – Ένα framework για πειράματα σε επεξεργασία φυσικής γλώσσας.
- **UIMA** – Ένα framework για ανάλυση μη δομημένου περιεχομένου όπως κείμενο, ήχος και βίντεο που αναπτύχθηκε αρχικά από την IBM.

2.3.1 WEKA

Το 1993 το Πανεπιστήμιο του Waikato στην Νέα Ζηλανδία ξεκίνησε την ανάπτυξη της αυθεντικής έκδοσης του Weka που έγινε ένα μείγμα από TCL/TK,c και Makefiles[13]. Το 1997 πάρθηκε η απόφαση να ξαναστηθεί το Weka εκ του μηδενός σε Java παρέχοντας εφαρμογές από modeling αλγορίθμους [10]. Το 2005 το Weka λαμβάνει το SIGKDD Data Mining and Knowledge Discovery Service Award[11][12]. Το 2006 η Pentaho Corporation εξασφάλισε μια αποκλειστική άδεια να χρησιμοποιεί το πρόγραμμα για business intelligence. Διαμόρφωσε το component για την εξόρυξη πληροφορίας και την ανάλυση προγνώσεων της πλατφόρμας Pentaho business intelligence[13].

Το WEKA (**W**aikato **E**nvironment for **K**nowledge **A**nalysis) είναι ένα open source πακέτο λογισμικού που υποστηρίζει την διαδικασία εξόρυξης πληροφορίας. Αναπτύχθηκε από το Πανεπιστήμιο του Waikato στην Νέα Ζηλανδία εξ ολοκλήρου στην Java. Παρέχει υποστήριξη για όλη την διαδικασία του πειραματικού data mining και χρησιμοποιείται στην εκπαίδευση, στην έρευνα αλλά και σε εφαρμογές. Προσφέρεται δωρεάν για download από την διεύθυνση http://www.cs.waikato.ac.nz/ml/weka/ και μπορεί να εγκατασταθεί σε διαφορετικά λειτουργικά συστήματα (Windows 32&64bit, Mac OS,Linux κα). Το WEKA διαθέτει πληθώρα από τεχνικές προ-επεξεργασίας(preprocessing tools), καθώς και αλγορίθμους για όλες τις κύριες εργασίες της διαδικασίας εξόρυξης πληροφορίας όπως είναι οι κατηγοριοποίηση (classification/regression), συσταδοποίηση (clustering) και η δημιουργία κανόνων συσχέτισης (association rules). Τέλος υποστηρίζει την παρουσίαση των δεδομένων και αποτελεσμάτων, πέρα από τα στατιστικά μέσα και με οπτικοποίηση (visualization).

Για τις ανάγκες της παρούσας εργασίας, έχει εγκατασταθεί η έκδοση 3.6 σε λειτουργικό Windows 7. Η εγκατάσταση είναι απλή χωρίς να παρουσιάζει καμία ιδιαιτερότητα, ενώ το αρχείο εγκατάστασης είναι 36,6MB. Συμπληρωματικά απαιτήθηκε η εγκατάσταση της Java VM 1.6.

Το WEKA διαθέτει 4 διαφορετικά interface:

Explorer: Είναι ίσως το πιο διαδεδομένο interface στους χρήστες του προγράμματος. Έχει όλες τις επιλογές μιας διαδικασίας, από την εισαγωγή του dataset ως και την οπτικοποίηση των αποτελεσμάτων, προσβάσιμες σε μορφή μενού απλοποιώντας έτσι κατά πολύ την διαδικασία για τον χρήστη. Μειονέκτημα του συγκεκριμένου interface είναι ότι σε πολύ μεγάλα μεγέθη δεδομένων αδυνατεί να γίνει επεξεργασία λόγω του ότι όλες οι πληροφορίες κρατούνται στην κύρια μνήμη κάνοντας έτσι μη αποδοτικό.

Experimenter: Αυτό το interface επιτρέπει την διαχείριση μεγάλης κλίμακας πειραμάτων, να τα τρέξεις να τα αφήσεις, και μετά να επιστρέψεις όταν τελειώσουν και να αναλύσεις τα στατιστικά απόδοσης που προέκυψαν. Αυτοματοποιεί την διαδικασία του πειράματος και υπάρχει ακόμα και η δυνατότητα να μοιραστεί ο φόρτος επεξεργασίας σε πολλαπλά μηχανήματα μέσω της Java RMI. Πολλές φορές χρησιμοποιείται αλληλεπιδραστικά με τον Explorer.

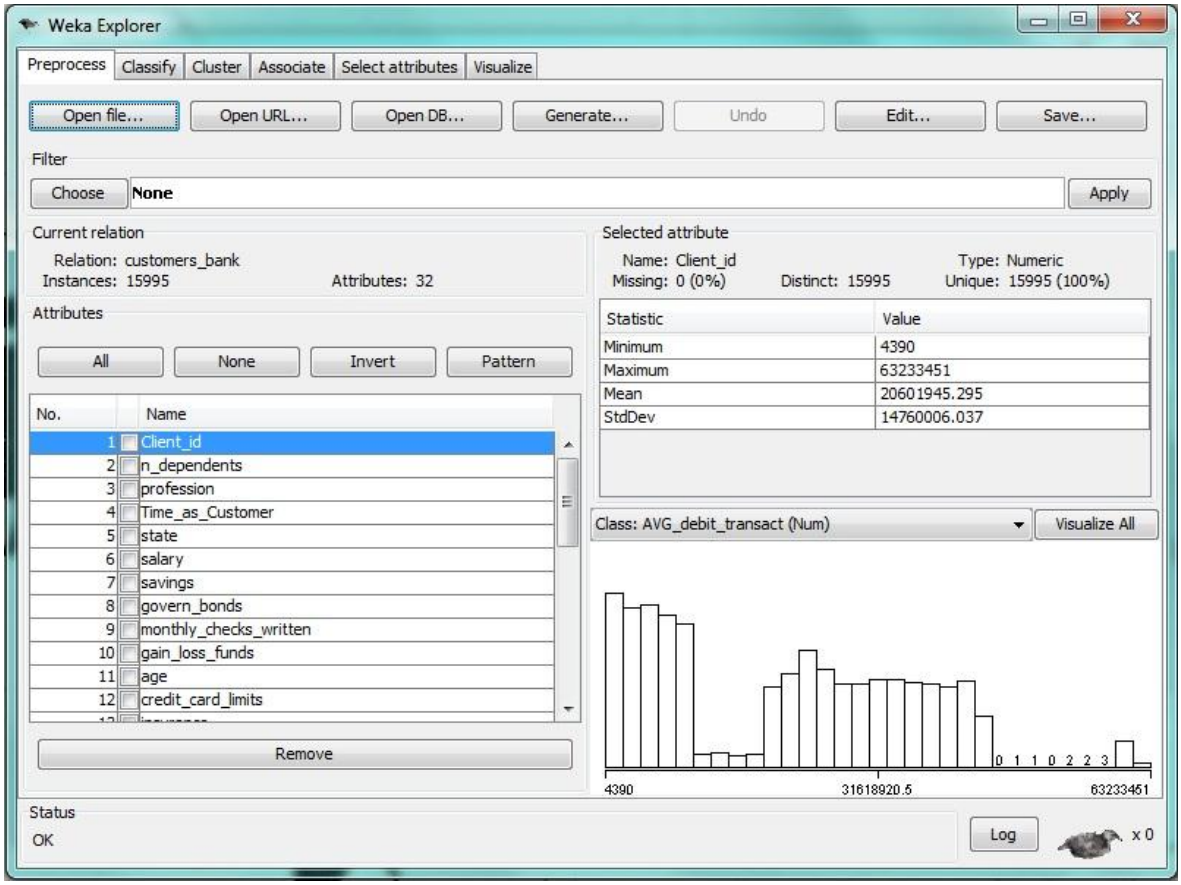
KnowledgeFlow: Είναι ακόμη ένα γραφικό interface όπως και τα προηγούμενα. Στο συγκεκριμένο διαφέρει το layout καθώς η διαδικασία «σχεδιάζεται» με γραμμικό τρόπο με την μορφή κουτιών που εισάγονται όπως είναι τα datasources και μέσω ενός ή περισσοτέρων γραμμών-μονοπατιών φαίνεται πως είναι δομημένη η διαδικασία από την αρχή ως το τέλος.

Simple CLI: Πρόκειται για το command line του προγράμματος όπου πρέπει κάποιος να έχει προγραμματιστικές γνώσεις για να το χρησιμοποιήσει. Στην παρούσα εργασία δεν θα ασχοληθούμε με αυτό το interface.

Στην εικόνα 2.1 βλέπουμε μια άποψη από το περιβάλλον του Weka στο Explorer interface.

Όπως αναφέρθηκε και παραπάνω, το Weka υποστηρίζει αρκετές διεργασίες για την εξόρυξη πληροφορίας, με βασικότερες τις clustering, classification και association rules. Όλες οι τεχνικές του Weka στηρίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα σε ένα αρχείο τύπου .csv ή τύπου .arff. Ο δεύτερος είναι ο τύπος αρχείων που δημιουργεί το WEKA. Σε κάθε περίπτωση, κάθε μία εγγραφή δεδομένων συνίσταται από τις τιμές ενός πλήθους χαρακτηριστικών, τα οποία μπορεί να είναι είτε ποσοτικά είτε ποιοτικά. Το πρόγραμμα προσφέρει πρόσβαση σε βάσεις δεδομένων μέσω SQL με την χρήση της σύνδεσης μέσω Java (Java Database Connectivity) και μπορεί να

επεξεργαστεί αποτελέσματα που προέρχονται από κάποιο ερώτημα (query) σε βάση δεδομένων [13].



Εικόνα 2.1: Το Explorer interface του Weka

Στην ιστοσελίδα του προγράμματος http://www.cs.waikato.ac.nz/ml/weka/index.html μπορούν να αντληθούν τα εγχειρίδια χρήσης (ανά έκδοση του προγράμματος), παρουσιάσεις σε powerpoint για όλα τα γραφικά περιβάλλοντα που έχει, οδηγό χρήσης για το interface explorer, και tutorials για το KnowledgeFlow και τον Experimenter.

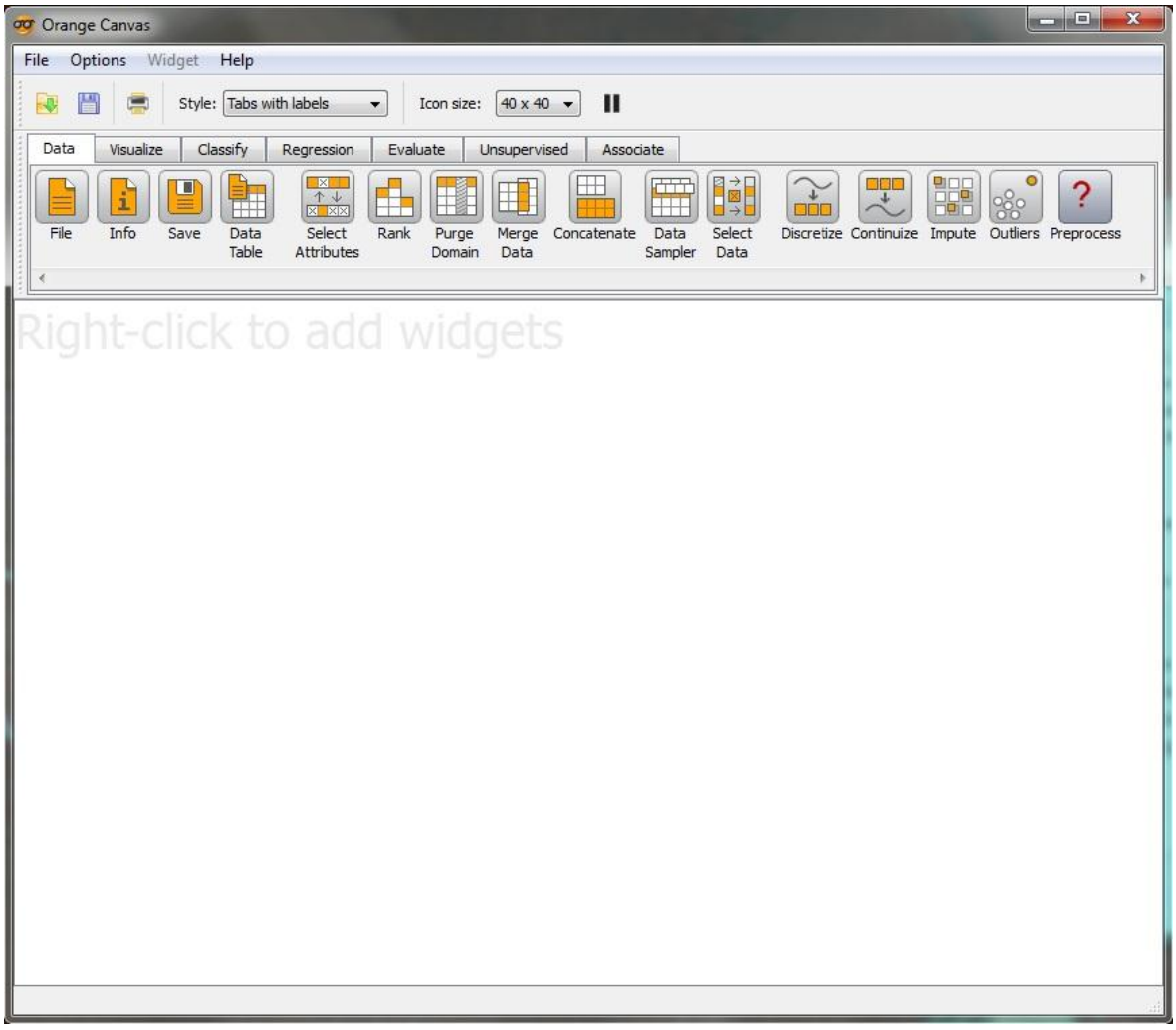
2.3.2 ORANGE

Το 1996, το Πανεπιστήμια της Λιουμπλιάνα και το Ινστιτούτο Jožef Stefan ξεκίνησαν την ανάπτυξη του ML που είναι ένα machine learning framework για την C++. Το 1997 «συνδέσεις» με την Python αναπτύχθηκαν για την ML ,που μαζί με τα modules της Python σχημάτισαν ένα ενιαίο framework που ονομάστηκε **Orange.** Κατά τα επόμενα χρόνια οι περισσότεροι σημαντικοί αλγόριθμοι για εξόρυξη πληροφορίας και machine learning προγραμματίστηκαν είτε σε C++ - που είναι ο πυρήνας του Orange - είτε σε Python modules. Το 2002 τα πρώτα πρωτότυπα για την δημιουργία ενός ευέλικτου γραφικού περιβάλλοντος σχεδιάστηκαν με την χρήση του Pmw Python megawidgets. Το 2003, το γραφικό περιβάλλον ξανασχεδιάστηκε και επαναπρογραμματίστηκε για το Qt framework, με την χρήση των PvQt Python «συνδέσεων»(bindings). Το οπτικό προγραμματιστικό framework ορίστηκε και η ανάπτυξη των widgets(γραφικά στοιχεία της pipeline ανάλυσης δεδομένων) ξεκίνησε. Το 2005 ,οι επεκτάσεις για ανάλυση δεδομένων στην βιοπληροφορική δημιουργήθηκαν. Το 2008 τα Mac OS X DMG και Fink-based αρχεία εγκατάστασης αναπτύχθηκαν. Το 2009 πάνω από 100 widgets δημιουργήθηκαν και διατηρήθηκαν. Από το 2009, το Orange είναι στην 2.0 beta έκδοση και το site προσφέρει αρχεία εγκατάστασης βασισμένα σε καθημερινό compilation cycle.[14]

Το Orange είναι ένα εργαλείο που αποτελείται από διάφορες συνιστώσες (components) πάνω στην εξόρυξη λογισμικού και παρέχει φιλικό περιβάλλον αλλά και ευέλικτο visual προγραμματισμό, και χρησιμοποιείται σε αναλύσεις δεδομένων και την οπτικοποίηση τους. Αναπτύσσεται από το Πανεπιστήμιο της Λιουμπλιάνα. Χρησιμοποιεί την γλώσσα προγραμματισμού Python και τις βιβλιοθήκες της για την δημιουργία script. Εμπεριέχει εκτενή σύνολα από στοιχεία για προ-επεξεργασία δεδομένων και φυσικά έχει και αλγορίθμους για τα βασικά μοντέλα και τεχνικές (clustering, classification κτλ). Υλοποιείται σε C++ που της δίνει το ατού της ταχύτητας, και σε Python που της δίνει και την ευελιξία. Είναι δωρεάν για download και το βρίσκουμε στην ιστοσελίδα http://orange.biolab.si/. Μπορεί να εγκατασταθεί σε 3 διαφορετικά λειτουργικά δηλαδή σε Windows, Mac OS X, και

Linux. Χρειάζεται να εγκατασταθούν μαζί με το πρόγραμμα και οι βιβλιοθήκες Python (Python, PythonWin, NumPy, PyQt, PyQwt). Για τις ανάγκες της παρούσας εργασίας, έχει εγκατασταθεί η έκδοση 2.0b, ενώ το αρχείο εγκατάστασης που τα εμπεριέχει είναι 58,4MB.

Στο κομμάτι του Documentation στο http://orange.biolab.si/doc/ του προγράμματος περιέχει κατάλογο όλων των widgets με περιγραφές και αναλύει με παραδείγματα που και πως χρησιμοποιείται το κάθε ένα. Επίσης έχει manual για τον προγραμματισμό Widget στο Orange και οδηγό εγκατάστασης γι αυτούς που έχουνε Mac OS. Πολύ σημαντικό είναι και το forum που διατηρείται με ανακοινώσεις για τις νέες εκδόσεις και συζητήσεις που μπορούν να γίνουν γύρω από το πρόγραμμα ακόμα και support για bugs.



Εικόνα 2.2: Μια άποψη από το περιβάλλον του Orange

2.3.3 Rapidminer

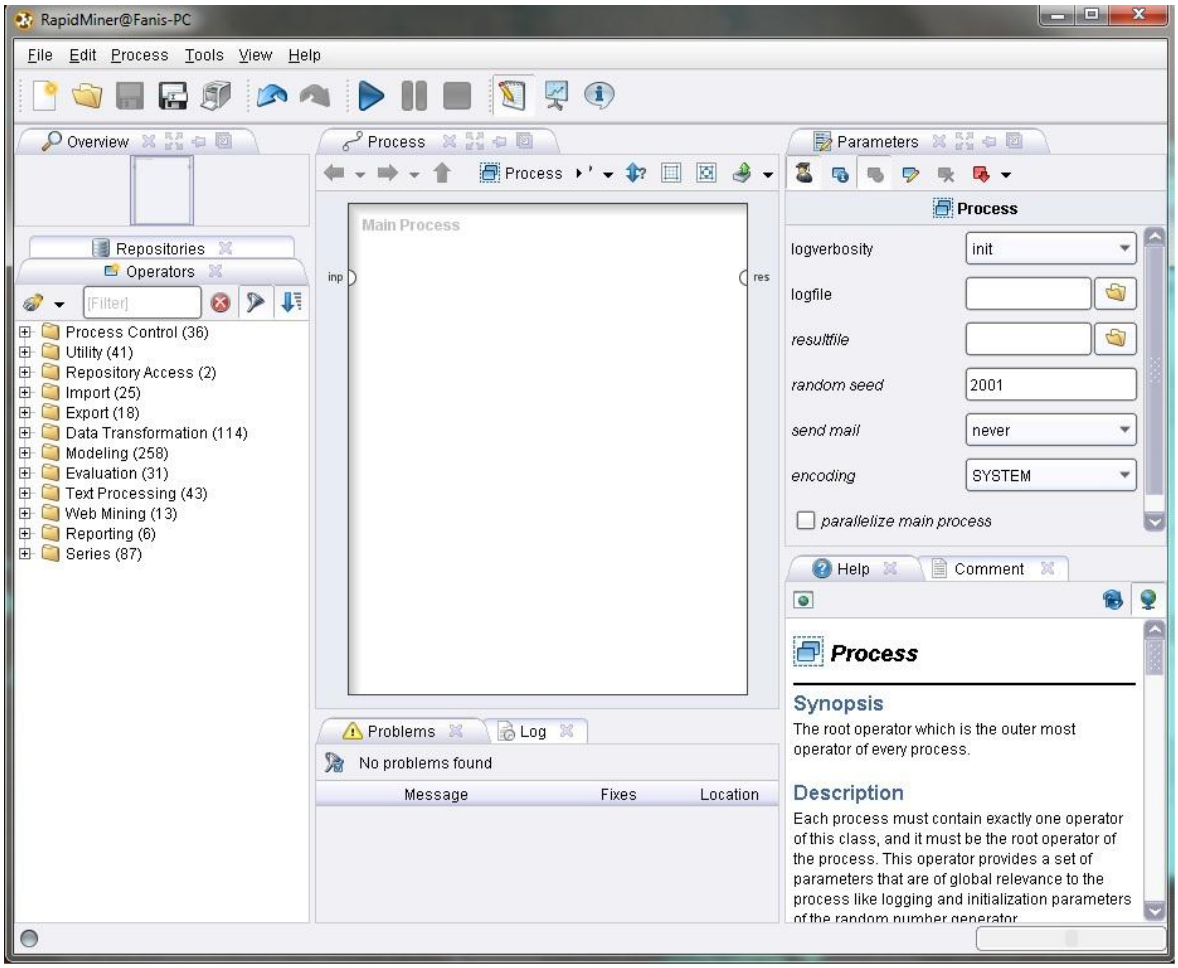
Το RapidMiner ξεκίνησε το 2001 από τους Ralf Klinkenberg, Ingo Mierswa και τον Simon Fischer στην μονάδα Τεχνητής Νοημοσύνης του Πανεπιστημίου του Ντόρτμουντ. Το 2006 ο Ingo Mierswa και ο Ralf Klinkenberg ιδρύσαν την εταιρεία Rapid-I που τώρα είναι ο βασικός συνεισφέρων στην περαιτέρω ανάπτυξη του προγράμματος όπως επίσης και περισσότεροι από 30 προγραμματιστές ανά τον κόσμο[9]. Είναι γραμμένο στην γλώσσα προγραμματισμού Java.

Το RapidMiner, παλαιότερα γνωστό και ως YALE (Yet Another Learning Environment), είναι ένα περιβάλλον για data mining, text mining, και άλλες χρήσεις. Χρησιμοποιείται για έρευνα, εκπαίδευση, εξάσκηση, για ανάπτυξη εφαρμογών και εφαρμογές για την βιομηχανία. Σε ψηφοφορία που διεξήχθη στην ιστοσελίδα για την εξόρυξη πληροφορίας, το RapidMiner κατέλαβε την δεύτερη θέση για εξόρυξη πληροφορίας και για εργαλεία ανάλυσης για πραγματικά project το 2009 [7] και κατέλαβε την πρώτη θέση το 2010 [8].

Υπάρχουν δυο εκδόσεις του προγράμματος:

Α) Η **Community edition** που είναι open source και διατίθεται δωρεάν στην ιστοσελίδα του προγράμματος http://rapid-i.com/content/view/26/84/lang,en/

Β) Η **Enterprise edition** η οποία έχει την Community Edition μαζί με Services και Guarantees και δεν είναι δωρεάν.



Εικόνα 2.3: Μια άποψη από το περιβάλλον του Rapidminer

Στην παρούσα εργασία χρησιμοποιείται η community edition και εγκαταστάθηκε η έκδοση 5.1.006 για windows 64bit. Υπάρχουν επίσης και οι εκδόσεις για windows 32 bit αλλά και Other Systems χωρίς όμως να κατονομάζει συγκεκριμένα λογισμικά. Το αρχείο της εγκατάστασης καταλαμβάνει χωρητικότητα 49,7MB και δεν υπάρχει κάποια άλλη απαίτηση εγκατάστασης λογισμικού.

Το RapidMiner υποστηρίζει και κάποιες επεκτάσεις (extensions) που κατά την εγκατάσταση μπορείς να εισάγεις στο πρόγραμμα. Παρατίθενται συνοπτικά:

- **R extension**: Υποστηρίζει πλέον την R console και όλα τα scripts που έχει.
- **Weka extension**: Μπορούν να χρησιμοποιηθούν όλα τα μοντέλα και όλες οι βιβλιοθήκες που υπάρχουν στο Weka.
- **Reporting extension**: Αναβαθμίζει τον τρόπο εμφάνισης των reports και μπορεί να εξάγει αναφορές ακόμη και σε pdf.
- **Text extension**: Εισάγονται όλοι οι operators για στατιστική ανάλυση στα κείμενα.

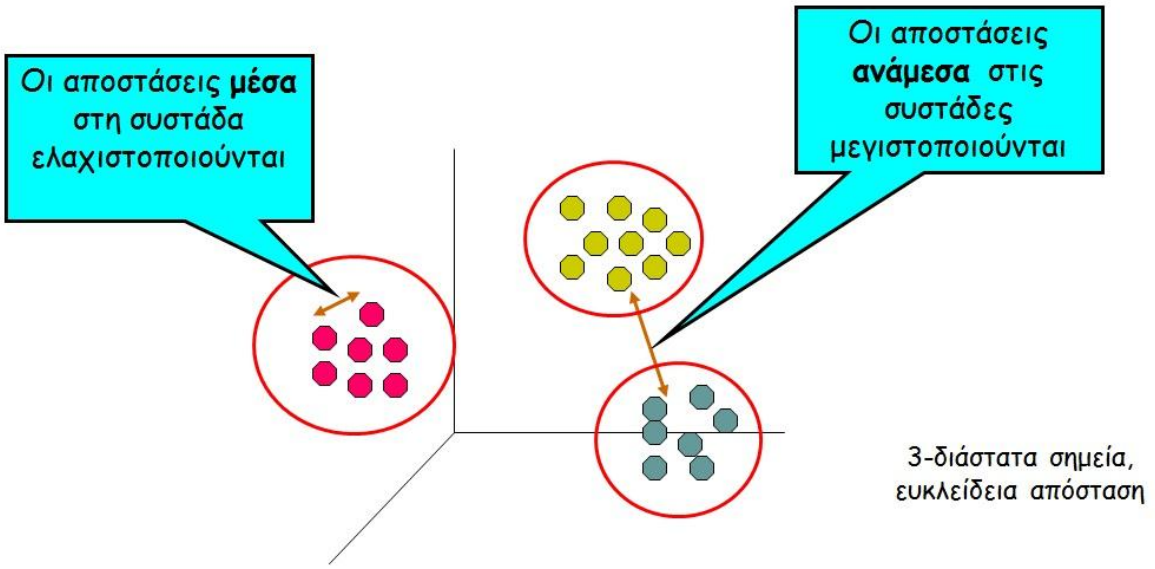
- **PaREn extension**: Είναι ένα αυτόματο σύστημα για classification που προτείνει αυτόματα ,κατασκευάζει, και βελτιστοποιεί μια διαδικασία classification.
- **PMML extension**: Παρέχει έναν ακόμα operator για δημιουργία μοντέλων σε PMML standard.
- **Web extension**: Παρέχει πρόσβαση σε web σελίδες και υπηρεσίες μαζί με operators για την διαχείρισή τους
- **Community extension**: Δίνει την δυνατότητα να μοιραστείς μέσω του myexperiment.org την πορεία δουλειάς και τα αποτελέσματα που προκύπτουν.
- **Series extension**: Βοηθάει σε περιπτώσεις σειριακών δεδομένων όπως οι χρονοσειρές.
- **Parallel Processing extension**: Εισάγει operators που βελτιώνουν την ταχύτητα αποτελεσμάτων σε ένα σύστημα με πολλούς πυρήνες στον επεξεργαστή.

Το RapidMiner στην ιστοσελίδα του προσφέρει Video tutorials,οδηγό εγκατάστασης και manual στα Αγγλικά και τα Γερμανικά. Έχει δική του σελίδα wiki, blog αλλά και πληρέστατο forum όπου μπορεί κάποιος να απευθυνθεί.

ΚΕΦΑΛΑΙΟ 3
Συσταδοποίηση (Clustering)

3.1 Εισαγωγή

Η συσταδοποίηση χωρίζει τα δεδομένα σε ομάδες(συστάδες) που έχουν νόημα ή είναι χρήσιμα ή και τα δυο. Αν ο στόχος είναι να βρούμε συστάδες που έχουν νόημα τότε οι συστάδες θα κρατάνε την φυσική δομή των δεδομένων. Ένας ορισμός για το τι είναι συσταδοποίηση είναι ο εξής: «Εύρεση συστάδων (ομάδων) αντικειμένων έτσι ώστε τα αντικείμενα σε κάθε συστάδα να είναι όμοια (ή να σχετίζονται) και διαφορετικά (ή μη σχετιζόμενα) από τα αντικείμενα των άλλων συστάδων»[15].



Εικόνα 3.1:Ένα παράδειγμα συστάδων σε τρεις διαστάσεις
Πηγή: http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf

Η συσταδοποίηση παίζει πολύ σημαντικό ρόλο σε διάφορους τομείς όπως είναι η ψυχολογία, βιολογία, στατιστική, αναγνώριση μοτίβων και φυσικά στην εξόρυξη πληροφορίας. Υπάρχουν διάφορα παραδείγματα που εμφανίζονται συστάδες όπως είναι πελάτες με όμοια συμπεριφορά, μετοχές με παρόμοια διακύμανση στην τιμή τους, σε ομαδοποίηση κειμένων, χαρακτηριστικά ασθενειών κα. Το clustering χρησιμοποιείται και για την κατανόηση των δεδομένων συνήθως με κάποιο εργαλείο εφαρμογή (όπως είναι αυτά που πραγματεύεται η εργασία) και

με την οπτικοποίηση των δεδομένων βγαίνουν κάποια συμπεράσματα για την κατανομή τους.

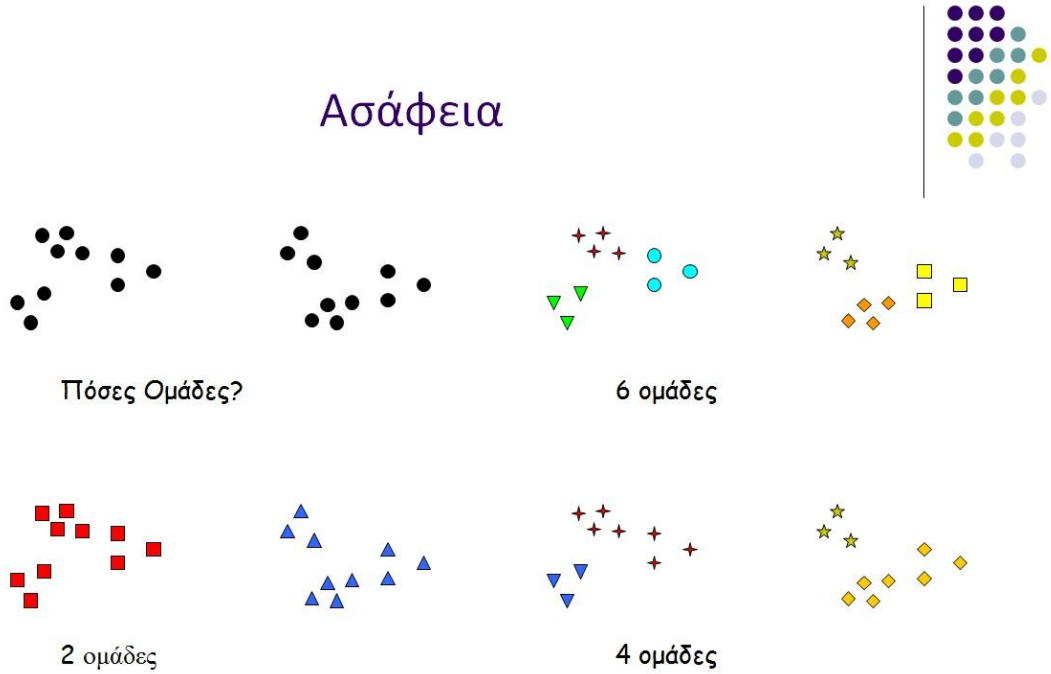
Περίληψη: Σε πρακτικό επίπεδο μερικές τεχνικές όπως είναι η παλινδρόμηση (regression) δεν συμφέρει να χρησιμοποιηθούν σε μεγάλα data sets, έτσι αντί να χρησιμοποιηθούν στον αλγόριθμο ολόκληρα τα data sets, μπορούν να μπούνε σε ένα μειωμένο αριθμό δεδομένων που θα αποτελούνται από τα **πρωτότυπα** των συστάδων που θα είναι αντιπροσωπευτικά σε σχέση με την συστάδα. Τα αποτελέσματα θα είναι παρόμοια με το αν χρησιμοποιούνταν ολόκληρα τα data sets αλλά με μεγαλύτερη ταχύτητα και με λιγότερη κατανάλωση πόρων του συστήματος.

Συμπίεση: Τα πρωτότυπα μπορούν επίσης να χρησιμοποιηθούν για συμπίεση των δεδομένων που είναι γνωστή ως vector quantization και συνήθως χρησιμοποιείται σε εικόνες ήχο ή βίντεο όταν α) πολλά από τα αντικείμενα-δεδομένα είναι παρόμοια το ένα με το άλλο, β) κάποια απώλεια από πληροφορία είναι αποδεκτή και γ) μια σημαντική μείωση του μεγέθους των δεδομένων είναι επιβεβλημένη.

Μπορεί να γίνει με την χρήση των πρωτοτύπων αποδοτική κατασκευή ευρετηρίων αλλά και εύρεση κοντινότερου γείτονα με μεγαλύτερη ταχύτητα και υψηλότερη αποδοτικότητα.

Πολλές φορές με τα δεδομένα που υπάρχουν υπάρχει ασάφεια ως προς την επιλογή η τον διαχωρισμό των συστάδων καθώς κάτι τέτοιο δεν είναι ευκρινές όπως βλέπουμε και στην εικόνα 3.2.

Προκύπτει το ερώτημα λοιπόν, μιας και δεν βοηθούν πάντα τα δεδομένα στο να έχουμε προφανείς συστάδες, πότε είναι η συσταδοποίηση καλή. «Μια μέθοδος συσταδοποίησης είναι καλή αν παράγει συστάδες καλής ποιότητας. Δηλαδή όταν υπάρχει **μεγάλη** ομοιότητα **εντός** της συστάδας και **μικρή** ομοιότητα ανάμεσα στις συστάδες. Σε κάθε περίπτωση η ποιότητα εξαρτάται από την μέτρηση ομοιότητας και από την μέθοδο υλοποίησης της συσταδοποίησης» [15].



Εικόνα 3.2: Παράδειγμα περίπτωσης με ασάφεια ως προς τις συστάδες
Πηγή: http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf

3.2 Κατηγορίες Μεθόδων

Ένα σύνολο από συστάδες αποτελούν την συσταδοποίηση και υπάρχουν διάφοροι τύποι συσταδοποίησης. Οι δυο βασικότεροι είναι οι: **Διαχωριστική συσταδοποίηση** (Partitional clustering) και η **Ιεραρχική Συσταδοποίηση** (Hierarchical clustering).

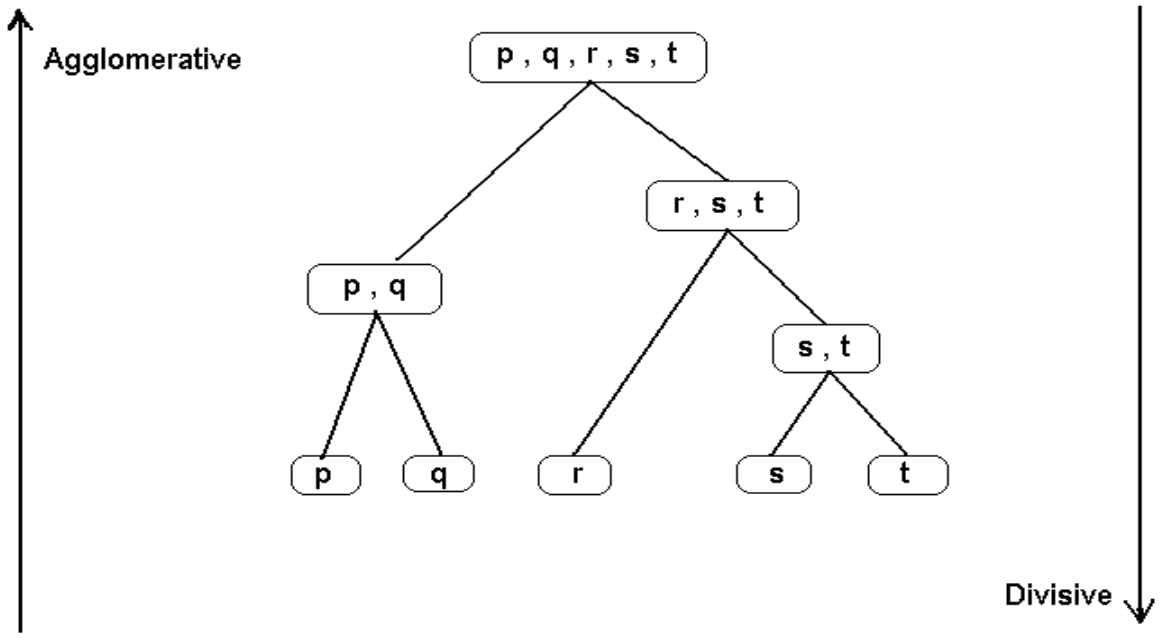
3.2.1 Hierarchical clustering

Στην Ιεραρχική Συσταδοποίηση τα δεδομένα δεν είναι χωρισμένα σε συγκεκριμένες συστάδες με ένα απλό βήμα. Αντίθετα γίνονται μια σειρά από διαχωρισμούς στους οποίους μπορεί όλα τα στοιχεία να μπούνε μόνο σε μια συστάδα ή ακόμα να προκύψει η αριθμός συστάδων που θα αποτελούνται από ένα μόνο στοιχείο. Με άλλα λόγια, η Ιεραρχική Συσταδοποίηση είναι ένα σύνολο από εμφωλευμένες συστάδες που είναι οργανωμένες υπό μορφή δέντρου. Κάθε κόμβος (συστάδα) στο δέντρο, εκτός από τους κόμβους-φύλλα, είναι η ένωση από τους κόμβους-παιδιά τους και η ρίζα του δέντρου είναι ο κόμβος που περιέχει όλα

τα στοιχεία. Συχνά, αλλά όχι πάντα, τα φύλλα του δέντρου είναι συστάδες που αποτελούνται από ένα και μόνο στοιχείο.

Υπάρχουν δυο βασικά πλεονεκτήματα της Ιεραρχικής Συσταδοποίησης. Το πρώτο είναι ότι δεν χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό από συστάδες. Οποιοσδήποτε επιθυμητός αριθμός από συστάδες μπορεί να επιτευχθεί κόβοντας το δενδρόγραμμα στο κατάλληλο επίπεδο. Τα δενδρογράμματα μιας ιεραρχίας μπορεί να αντιστοιχούνε σε λογικές αντιστοιχήσεις πχ στις βιολογικές επιστήμες (ζωικό βασίλειο, φυλογένεια κα) [16].

Η Ιεραρχική Συσταδοποίηση χωρίζεται στις **συσσωρευτικές (agglomerative)** μεθόδους που είναι μια σειρά από συγχωνεύσεις από n στοιχεία σε ομάδες, και στις μεθόδους **διαίρεσης(divisive)** που διαχωρίζουν n στοιχεία διαδοχικά σε μικρότερες ομάδες. Οι agglomerative μέθοδοι χρησιμοποιούνται πιο συχνά. Σε αυτές τις μεθόδους η διαδικασία ξεκινά παίρνοντας το κάθε σημείο ως ξεχωριστή συστάδα και σε κάθε βήμα συγχωνεύει το πιο κοντινό ζευγάρι συστάδων μέχρι να μείνει μόνο μια (ή k) συστάδα. Στις divisive μεθόδους η διαδικασία είναι αντίστροφη. Αρχίζει με μια συστάδα που περιέχει όλα τα σημεία και σε κάθε βήμα, διαχωρίζει μια συστάδα έως κάθε συστάδα να περιέχει μόνο ένα σημείο. Στο παρακάτω διάγραμμα (Εικόνα 3.3) ,γνωστό και ως δενδρόγραμμα φαίνονται οι συγχωνεύσεις και οι διαιρέσεις που γίνονται σε κάθε στάδιο της ανάλυσης.



Εικόνα 3.3: Φαίνεται η αντίστροφη διαδικασία ανάμεσα στις δυο υποομάδες της Ιεραρχικής Ομαδοποίησης

Πηγή: http://www.resample.com/xlminer/help/HClst/Clustering1.gif

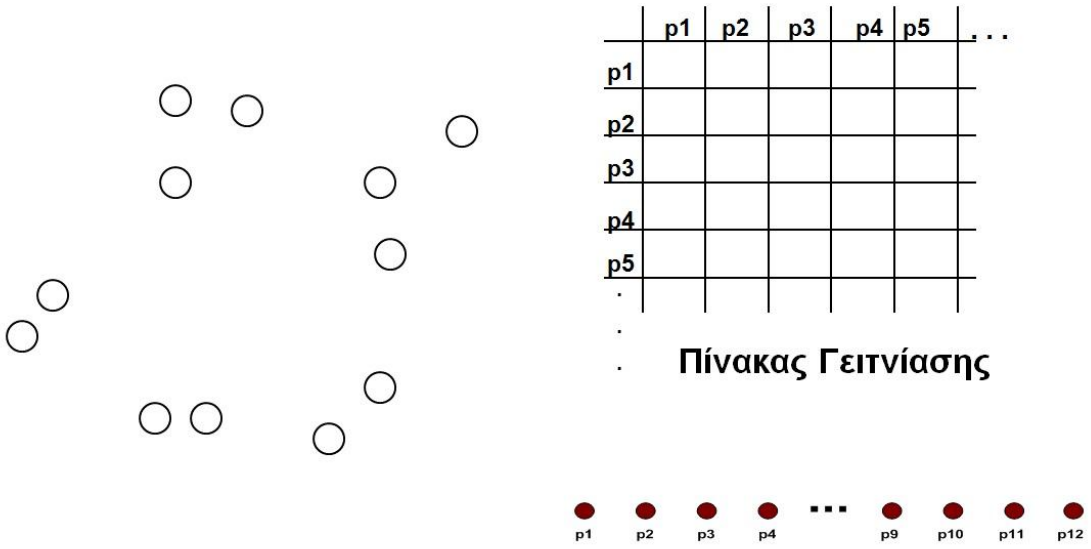
Οι παραδοσιακοί αλγόριθμοι για την Ιεραρχική συσταδοποίηση χρησιμοποιούν έναν πίνακα ομοιότητας ή απόστασης και γίνεται διαχωρισμός ή συγχώνευση (ανάλογα με το αν είναι divisive ή agglomerative) μιας ομάδας σε κάθε βήμα. Πολλοί συσσωρευτικοί αλγόριθμοι είναι παραλλαγές της εξής φιλοσοφίας: ξεκινώντας με μοναδικά στοιχεία ως αρχικές συστάδες, διαδοχικά ενώνουμε τις δυο κοντινότερες συστάδες μέχρι να μείνει μόνο μια.

Εδώ παρατίθεται υπό μορφή ψευδογλώσσας ο βασικός αλγόριθμος:

1: Υπολογισμός του Πίνακα Γειτνίασης
2: Έστω κάθε σημείο αποτελεί και μια συστάδα
3: **ΕΠΑΝΑΛΗΨΗ**
4: Συγχώνευση των δυο κοντινότερων συστάδων
5: Ενημέρωση του Πίνακα Γειτνίασης
6: **ΜΕΧΡΙ** να μείνει μία μόνο συστάδα

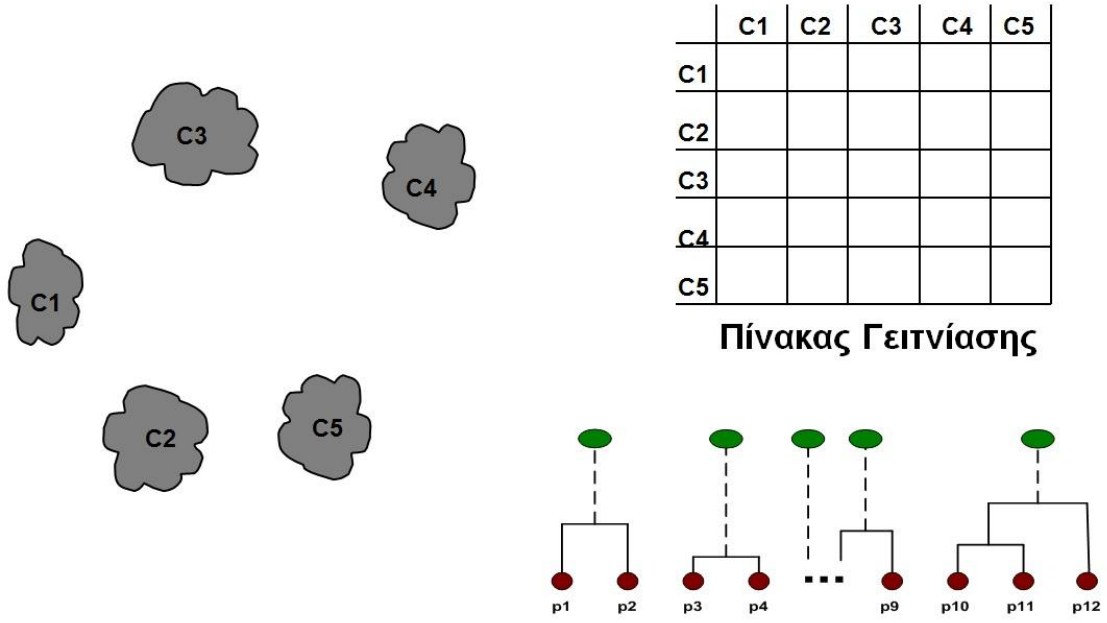
Όπως προκύπτει και από τον κώδικα, βασική λειτουργία είναι ο υπολογισμός γειτνίασης δυο συστάδων. Η διαφοροποίηση ανάμεσα τους διαφορετικούς αλγορίθμους έγκειται στο πως ορίζεται η απόσταση ανάμεσα σε δυο συστάδες. Ας δούμε όμως την διαδικασία με περισσότερη λεπτομέρεια.

Αρχικά έχουμε κάθε σημείο ως συστάδα και γίνεται ο πίνακας γειτνίασης.



Εικόνα 3.4: Το αρχικό σχήμα
Πηγή: http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf

Μετά από κάποιες συγχωνεύσεις έχουμε κάποιες συστάδες όπως στην Εικόνα 3.5



Εικόνα 3.5: Πως έχει γίνει μετά από κάποιες επαναλήψεις
Πηγή: http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf

Όσο υπήρχαν απλά στοιχεία ήταν προφανές πως θα υπολογιζόταν η απόσταση. Ο προβληματισμός προκύπτει όταν πρέπει να υπολογιστεί η απόσταση ανάμεσα σε δυο συστάδες που όπως είναι λογικό αποτελούνται από

πολλά σημεία. Υπάρχουν διάφορες τεχνικές και τρόποι για να μετρήσεις την απόσταση ανάμεσα σε 2 συστάδες που εξηγούνται παρακάτω.

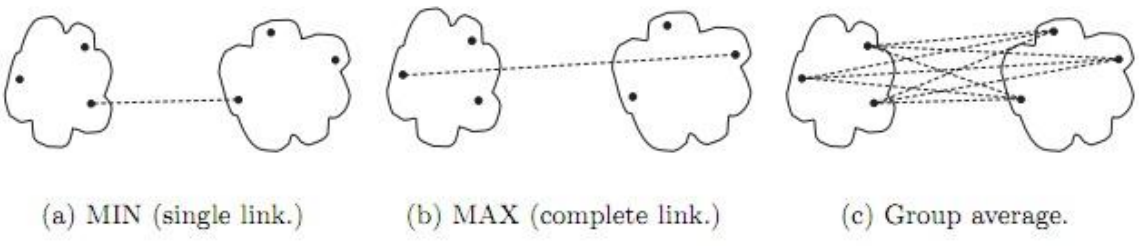
MIN ή μοναδικής ακμής ή μοναδικού συνδέσμου (single link). Η απόσταση μεταξύ δυο συστάδων βασίζεται στα δυο πιο **γειτονικά** σημεία στις διαφορετικές συστάδες. Ονομάζεται και μέθοδος συσταδοποίησης του κοντινότερου γείτονα(nearest neighbor). Το θετικό σημείο αυτού του τρόπου είναι ότι μπορεί να χειριστεί μη ελλειπτικά σχήματα και το αρνητικό ότι είναι ευαίσθητο σε θόρυβο και σε ακραίες τιμές(outliers).

MAX ή πλήρους συνδεσιμότητας (complete linkage). Η απόσταση μεταξύ δυο συστάδων βασίζεται στα δυο πιο **απομακρυσμένα** σημεία στις διαφορετικές συστάδες(longest edge). Υπολογίζει δηλαδή σε κάθε συστάδα την απόσταση παίρνοντας ως ακραία σημεία τα πιο απομακρυσμένα σε κάθε συσχέτιση. Όπως και στην MIN έτσι και στην MAX η επιλογή των σημείων καθορίζεται από όλα τα ζεύγη τιμών στις δυο συστάδες. Το θετικό σημείο είναι ότι υπάρχει λιγότερη εξάρτηση σε θόρυβο και σε ακραίες τιμές(outliers) και τα αρνητικά ότι τείνει να διασπά μεγάλες συστάδες και ότι οδηγεί συνήθως σε κυκλικά σχήματα.

Group Average ή αλλιώς μέσος όρος της ομάδας. Η εγγύτητα(proximity) δυο συστάδων είναι η μέση απόσταση μεταξύ όλων των ζευγαριών από τις δυο διαφορετικές συστάδες. Γίνεται η χρήση της μέσης απόστασης γιατί η ολική θα έδινε προτίμηση στις μεγάλες συστάδες. Συμβιβάζει τους δυο προηγούμενους τρόπους, και έχει το πλεονέκτημα ότι έχει μικρότερη ευαισθησία σε θόρυβο και ακραίες τιμές(outliers) και το μειονέκτημα ότι ευνοεί κυκλικές συστάδες.

$$\mathbf{proximity(Cluster_i, Cluster_j)} = \frac{\sum_{\substack{p_i \in Cluster_i \\ p_j \in Cluster_j}} \mathbf{proximity(p_i, p_j)}}{\mathbf{|Cluster_i| * |Cluster_j|}}$$

Στην εικόνα 3.6 φαίνεται σε σχήμα ο κάθε ένας από τους 3 τρόπους υπολογισμού της απόστασης ανάμεσα σε δυο συστάδες.



(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Εικόνα 3.6: Οι 3 τρόποι MIN,MAX, και Group Average

Πηγή: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Υπάρχουν κι άλλες μέθοδοι όπως είναι η απόσταση μεταξύ των κεντρικών σημείων δυο συστάδων ή η μέθοδος του Ward που χρησιμοποιεί τετραγωνικά λάθη.

Συμπερασματικά προκύπτει ότι όταν συγχωνευθούν δυο συστάδες αυτό μετά δεν αλλάζει και δεν ελαχιστοποιείται άμεσα κάποια αντικειμενική συνάρτηση. Διαφορετικές τεχνικές έχουν προβλήματα ένα η περισσότερα από τα παρακάτω: Ευαισθησία σε θόρυβο και outliers, δυσκολία στην διαχείριση συστάδων διαφορετικού μεγέθους ή κυρτού σχήματος ,και τέλος διάσπαση μεγάλων συστάδων.

3.2.2 Partitional Clustering

Οι αλγόριθμοι στην Διαχωριστική συσταδοποίηση δημιουργούν διάφορους διαχωρισμούς και μετά τους αξιολογούν με κάποιο κριτήριο. Συχνά αναφέρονται ως μη-Ιεραρχικοί(nonhierarchical) καθώς κάθε σημείο τοποθετείται σε ακριβώς ένα από k αποκλειστικές συστάδες. Επειδή μόνο ένα σετ από συστάδες είναι το αποτέλεσμα από έναν τυπικό αλγόριθμο διαχωριστικής συσταδοποίησης, ζητείται από τον χρήστη να θέσει τον επιθυμητό αριθμό από συστάδες. Ένας από τους πιο διαδεδομένους partitional αλγορίθμους είναι ο **k-means.** Ο χρήστης καλείται να παρέχει τον αριθμό των συστάδων (k) πριν ξεκινήσει και ο αλγόριθμος ορίζει πρώτα τα κέντρα από τα k μέρη. Εν ολίγοις, ο k-means μετά καταμερίζει τα σημεία βασιζόμενο στα τωρινά κέντρα και στην συνέχεια επανακαθορίζει τα κέντρα βασιζόμενο στα τωρινά μέλη της κάθε συστάδας. Αυτά τα 2 βήματα επαναλαμβάνονται μέχρι να βελτιστοποιηθεί ότι υπάρχει ομοιότητα στα στοιχεία εντός της κάθε συστάδας και διαφορετικότητα, μη-ομοιότητα στοιχείων από διαφορετικές συστάδες. Συνεπώς, η λογική εκκίνηση των κέντρων είναι ένας πολύ σημαντικός παράγοντας στο να πάρουμε σωστά αποτελέσματα από αλγορίθμους διαχωριστικής συσταδοποίησης [17]. Ένας ακόμα πολύ σημαντικός αλγόριθμος αυτής της κατηγορίας είναι ο k-medoids. Όπως και ο k-means προσπαθεί να ελαχιστοποιήσει το squared error, την απόσταση ανάμεσα σε σημεία που μπαίνουν σε μια συστάδα και ένα σημείο που έχει οριστεί ως το κέντρο της

συστάδας. Σε αντίθεση με τον k-means, ο k-medoids επιλέγει datapoints ως κέντρα.

Περισσότερα ως προς ψευδοκώδικα και για την εφαρμογή του partional clustering θα υπάρχουν σε επόμενη ενότητα όπου αναπτύσσεται ο αλγόριθμος k-means.

3.2.3 Άλλες μέθοδοι

Υπάρχουν οι **Density-based** αλγόριθμοι που επινοήθηκαν για να ανακαλύπτουν τυχαίου σχήματος συστάδες. Με αυτή την προσέγγιση, μια συστάδα θεωρείται ως μια περιοχή στην οποία τα αντικείμενα δεδομένων ξεπερνούν ένα όριο. Δυο βασικοί αλγόριθμοι αυτού του είδους είναι οι DBSCAN και OPTICS.

Οι **Subspace** αλγόριθμοι ψάχνουν για συστάδες που μπορούν να βρεθούν μόνο σε μια συγκεκριμένη προβολή (subspace,manifold) των δεδομένων. Αυτοί οι αλγόριθμοι άρα μπορούν να αγνοήσουν άσχετα χαρακτηριστικά. Το γενικό πρόβλημα είναι γνωστό ως **Correlation (αντιστοιχία) clustering** ενώ η ειδική περίπτωση των παράλληλων στους άξονες subspaces είναι επίσης γνωστή ως Two-way clustering ,co-clustering ή biclustering. Σε αυτές τις μεθόδους δεν μπαίνουν στις συστάδες μόνο τα αντικείμενα αλλά και χαρακτηριστικά των αντικειμένων και αν τα δεδομένα είναι της μορφής πίνακα τότε οι γραμμές και οι στήλες μπαίνουν σε συστάδες ταυτόχρονα. Συνήθως δεν δουλεύουν με τυχαίους συνδυασμούς χαρακτηριστικών όπως στους γενικούς αλγορίθμους subspace.[18] Δυο αντιπροσωπευτικοί αλγόριθμοι αυτής της κατηγορίας είναι ο CLIQUE(Agrawal et al. 2005) και ο SUBCLU(Kailing, Kriegel & Kröger 2004).

Μια ακόμα βασική ομάδα αλγορίθμων είναι οι **Distribution-based**. Το μοντέλο συσταδοποίησης που είναι πιο άμεσα συσχετισμένο με στατιστικά βασίζεται στα μοντέλα κατανομής (distribution models). Οι συστάδες μπορούν εύκολα να καθοριστούν ως αντικείμενα που ανήκουν πιο πιθανά στην ίδια διανομή. Μια καλή ιδιότητα αυτής της προσέγγισης είναι ότι μοιάζει πολύ με τον τρόπο που δημιουργούνται τα τεχνητά (artificial) δεδομένα, δειγματίζοντας δηλαδή τυχαία αντικείμενα μιας διανομής. Ενώ το θεωρητικό θεμέλιο αυτών των μεθόδων είναι εξαιρετικό, έχουν ένα πολύ σημαντικό πρόβλημα γνωστό ως overfitting, εκτός

αν μπουν περιορισμοί στην πολυπλοκότητα του μοντέλου. Ένα πιο πολύπλοκο μοντέλο συνήθως θα μπορεί να εξηγήσει τα δεδομένα καλύτερα, που κάνει την επιλογή της κατάλληλης πολυπλοκότητας μοντέλου δύσκολη. Η πιο διαπρεπής μέθοδος είναι γνωστή ως αλγόριθμος expectation-maximization (**EM-clustering**). Εδώ, το dataset είναι συνήθως φτιαγμένο με συγκεκριμένο αριθμό από διανομές Gaussian (για να αποφευχθεί το overfitting), που αρχικοποιούνται τυχαία και των οποίων οι παράμετροι είναι βελτιστοποιημένοι να ταιριάζουν καλύτερα στο dataset.

3.3 K-means

Ο K-means είναι ο βασικότερος και πιο διαδεδομένος αλγόριθμος που χρησιμοποιείται από τους διαχωριστικούς (partitional). Είναι βασισμένος στην έννοια της ομοιότητας / εγγύτητας και το χαρακτηριστικό του γνώρισμα είναι ότι κάθε συστάδα συσχετίζεται με ένα κεντρικό σημείο (centroid). Κάθε σημείο ανατίθεται στην συστάδα με το κοντινότερο κεντρικό σημείο. Βασικό στοιχείο επίσης του αλγορίθμου είναι ότι ο χρήστης προκαθορίζει τον αριθμό των συστάδων που θα κατασκευαστούν.

3.3.1 Περιγραφή

Ο τρόπος που λειτουργεί ο k-means είναι απλός και περιγράφεται στην συνέχεια. Αρχικά επιλέγεται ένας αριθμός k αρχικών κέντρων, όπου k είναι μια καθορισμένη απ τον χρήστη παράμετρος που είναι ο αριθμός των συστάδων που επιθυμεί. Στην συνέχεια κάθε σημείο ανατίθεται στο κοντινότερο του κέντρο και κάθε σύνολο σημείων που έχουν ανατεθεί σε ένα κέντρο αποτελούν μια συστάδα. Το κέντρο, ακολούθως, κάθε συστάδας επανακαθορίζεται βασιζόμενο στα σημεία που έχουν ανατεθεί στην συστάδα. Αυτό επαναλαμβάνεται μέχρι κανένα σημείο να μην αλλάξει συστάδα ή με άλλα λόγια κανένα κέντρο να μην μετακινηθεί [16]. Το κεντρικό σημείο είναι συνήθως ο μέσος (mean) των σημείων της συστάδας, ο οποίος μπορεί να μην είναι ένα από τα δεδομένα εισόδου. Παρακάτω ακολουθεί ο ψευδοκώδικας του αλγορίθμου:

1: Επιλογή *K* σημείων ως τα αρχικά κεντρικά σημεία

2: Repeat

3: Ανάθεση όλων των αρχικών σημείων στο **κοντινότερο** τους από τα *K* κεντρικά σημεία

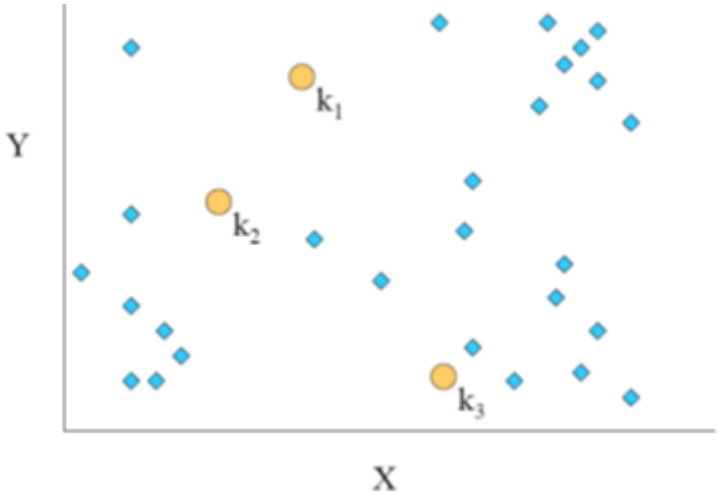
4: Επανα-υπολογισμός του **κεντρικού σημείου** κάθε συστάδας

5: Until τα κεντρικά σημεία να μην αλλάζουν

Για να ανατεθεί ένα σημείο στο κοντινότερο κέντρο, χρειάζεται ένα τρόπο μέτρησης της εγγύτητας που να ποσοτικοποιεί την έννοια του «κοντινότερου» για συγκεκριμένα δεδομένα. Η Ευκλείδεια απόσταση χρησιμοποιείται συνήθως για σημεία δεδομένων, ενώ η ομοιότητα συνημίτονου είναι πιο αρμόζουσα για έγγραφα. Όμως μπορεί να υπάρχουν διάφοροι τύποι μέτρησης της εγγύτητας που να ταιριάζουν για συγκεκριμένα τύπο δεδομένων. Για παράδειγμα η απόσταση Manhattan μπορεί να χρησιμοποιηθεί για αριθμητικά δεδομένα, ενώ η μέτρηση Jaccard μπορεί να χρησιμοποιηθεί για έγγραφα.

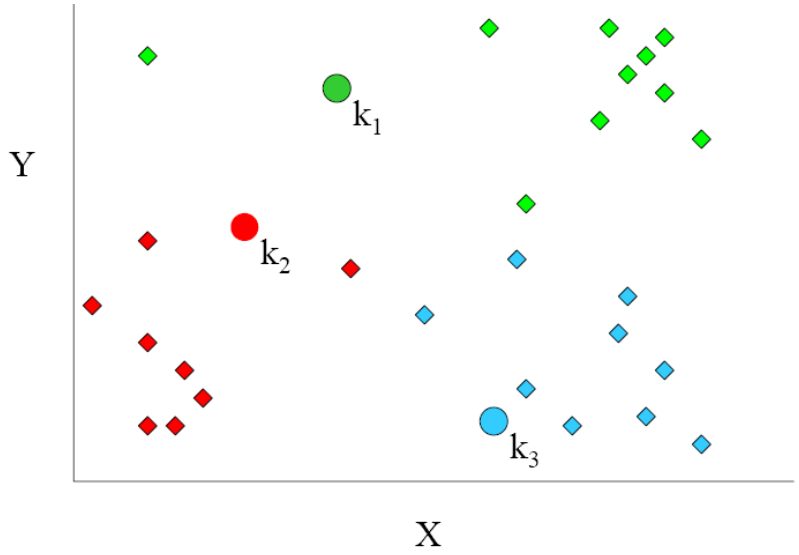
Συνήθως οι τρόποι μέτρησης που χρησιμοποιούνται για τον K-means είναι σχετικά απλοί από την στιγμή που ο αλγόριθμος επαναλαμβανόμενα υπολογίζει την απόσταση κάθε σημείου από το κάθε κέντρο [16].

Ας δούμε τώρα πως λειτουργεί ο αλγόριθμος αναλυτικά. Για παράδειγμα, ορίζουμε ότι θέλουμε 3 συστάδες και βλέπουμε πως μπαίνουν τυχαία τα τρία κεντρικά σημεία στην Εικόνα 3.7



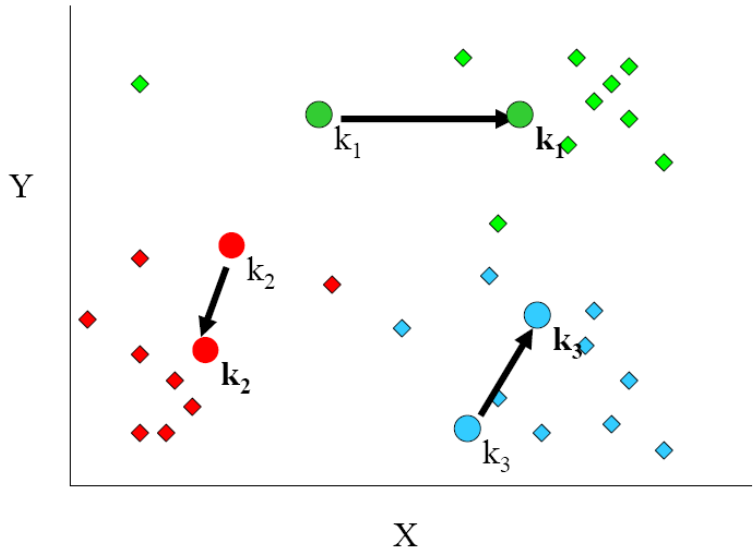
Εικόνα 3.7:Αρχική κατάσταση
Πηγή: http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf

Στην συνέχεια ανατίθενται τα σημεία στις 3 συστάδες σε σχέση με το πιο κοντινό σημείο. (Εικόνα 3.8)



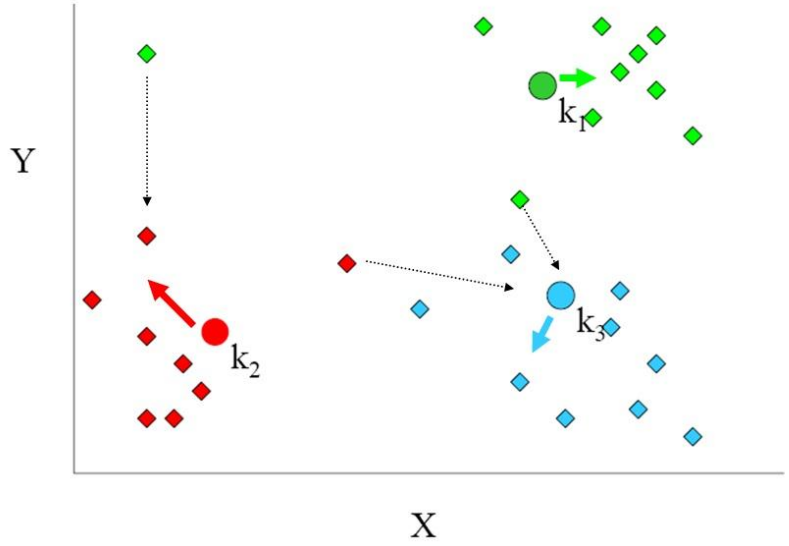
Εικόνα 3.8:Ανάθεση σημείων στις συστάδες
Πηγή: http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf

Μετά γίνεται επανα-υπολογισμός του κέντρου κάθε συστάδας όπως φαίνεται στην Εικόνα 3.9



Εικόνα 3.9:Επανα-υπολογισμός των κέντρων των συστάδων
Πηγή: http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf

Γίνεται νέα ανάθεση των σημείων με βάση τα νέα κέντρα (Εικόνα 3.10)



Εικόνα 3.10: Επανα-υπολογισμός των κέντρων των συστάδων
Πηγή: ***http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf***

Και στην συγκεκριμένη περίπτωση των εικόνων σταματάει ο αλγόριθμος εδώ γιατί στο επόμενο βήμα τα κέντρα δεν μετακινήθηκαν άλλο.

3.3.2 Αξιολόγηση

Το 4^{ο} βήμα του βασικού αλγορίθμου του k-means δηλώθηκε ως «Επανα-υπολογισμός του **κεντρικού σημείου** κάθε συστάδας». Με αυτόν τον τρόπο, το κέντρο μπορεί να διαφέρει εξαρτώμενο από το μέτρο της εγγύτητας των δεδομένων. Η ποιότητα μίας συσταδοποίησης μπορεί να «μετρηθεί» από μια αντικειμενική συνάρτηση, η οποία λαμβάνει υπόψη τις εγγύτητες των σημείων είτε μεταξύ τους είτε σε σχέση με το κέντρο, όπως είναι η ελαχιστοποίηση της απόστασης από κάθε σημείο στο κοντινότερό του κέντρο. Όμως η κεντρική ιδέα είναι η εξής: όταν έχουμε καθορίσει ένα μέτρο εγγύτητας και μια συνάρτηση, το κέντρο επιλέγουμε καθορίζεται μαθηματικά.

Ας θεωρήσουμε ότι έχουμε επιλέξει ως μέτρο εγγύτητάς την Ευκλείδεια Απόσταση. Για την αντικειμενική μας συνάρτηση, που μετράει την ποιότητα μιας συσταδοποίησης, χρησιμοποιούμε το **άθροισμα τετραγωνικού λάθους (Sum of the Squared Error-SSE)**, το οποίο εκφράζει διασπορά. Με άλλα λόγια, υπολογίζουμε το λάθος καθενός από τα σημεία, στην συγκεκριμένη περίπτωση την Ευκλείδεια απόσταση από το κοντινότερό του κέντρο, και μετά υπολογίζουμε το γενικό άθροισμα των τετραγωνικών λαθών. Από τα δυο διαφορετικά σύνολα

συστάδων που θα προκύψουν από δυο διαφορετικές εκτελέσεις του k-means, προτιμούμε αυτή με το μικρότερο τετραγωνικό λάθος από την στιγμή που αυτό σημαίνει πως τα κέντρα αυτής της συσταδοποίησης αντιπροσωπεύουν καλύτερα τα σημεία σε κάθε συστάδα. Εδώ έχουμε την συνάρτηση:

Πινακας 3.1
Πηγή: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Symbol	Description
$\mathbf{x}$	An object.
C_i	The i^{th} cluster.
$\mathbf{c}_i$	The centroid of cluster C_i.
$\mathbf{c}$	The centroid of all points.
m_i	The number of objects in the i^{th} cluster.
m	The number of objects in the data set.
K	The number of clusters.

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

Όπου *dist* είναι η Ευκλείδεια (L_2) απόσταση ανάμεσα σε δυο αντικείμενα στον Ευκλείδειο χώρο. Με αυτό τον τύπο αποδεικνύεται ότι το κέντρο που ελαχιστοποιεί το SSE της συστάδας είναι ο μέσος(mean). Το κέντρο-μέσος της i συστάδας ορίζεται από την παρακάτω συνάρτηση:

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

Για παράδειγμα, το κέντρο μιας συστάδας που περιέχει τα 3 σημεία που βρίσκονται σε δυο διαστάσεις (1,1), (2,3) και (6,2) είναι ((1 + 2 + 6) / 3),((1 + 3 + 2) / 3)=(3,2).

Τα βήματα 3 και 4 του αλγορίθμου k-means απευθείας προσπαθούν να ελαχιστοποιήσουν το SSE. Όμως οι ενέργειες του k-means στα βήματα 3 και 4 βρίσκουν μόνο το τοπικό ελάχιστο με το SSE καθώς βασίζονται στο να βελτιώσουν το SSE για συγκεκριμένες επιλογές κέντρων και συστάδων, από το να έχουν όλες τις πιθανές επιλογές.

Για να απεικονίσουμε ότι ο k-means δεν περιορίζεται σε δεδομένα στον Ευκλείδειο χώρο, παίρνουμε ένα άλλο παράδειγμα όπως είναι τα δεδομένα εγγράφου και το μέτρο ομοιότητας συνημιτόνου. Στόχος μας είναι να

μεγιστοποιήσουμε την ομοιότητα των εγγράφων στην συστάδα στο κέντρο της συστάδας. Αυτή η ποσότητα είναι γνωστή ως **συνοχή (cohesion)** της συστάδας. Για αυτό τον σκοπό μπορεί να δειχτεί ότι το κέντρο της συστάδας είναι όπως και στα Ευκλείδεια δεδομένα, ο μέσος. Η ανάλογη ποσότητα του ολικού SSE είναι η ολική συνοχή(cohesion), που δίνεται από την παρακάτω συνάρτηση:

$$\text{Total Cohesion} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} cosine(\mathbf{x}, \mathbf{c}_i)$$

Υπάρχει ένας αριθμός επιλογών για τον υπολογισμό της εγγύτητας, των κέντρων, και των αντικειμενικών συναρτήσεων που μπορούν να χρησιμοποιηθούν στον βασικό αλγόριθμο k-means και εγγυώνται την σύγκλιση. Ο πίνακας 3.2 δείχνει κάποιες πιθανές επιλογές συμπεριλαμβανομένων και των δυο που έχουμε ήδη συζητήσει.

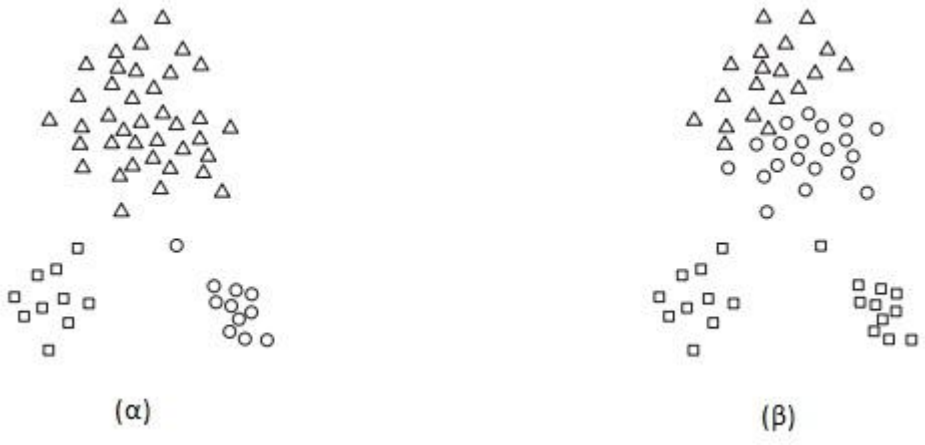
Πίνακας 3.2: Κάποιες βασικές συναρτήσεις υπολογισμού της εγγύτητας
Πηγή: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Συνάρτηση εγγύτητας	Κέντρο	Συνάρτηση αντικειμένου
Manhattan (L_1)	Διάμεσος	Ελαχιστοποίηση αθροίσματος της L1 απόστασης από ένα αντικείμενο με το κέντρο της συστάδας
Squared Euclidean (L_2^2)	Μέσος	Ελαχιστοποίηση αθροίσματος της απόστασης L2 από ένα αντικείμενο με το κέντρο της συστάδας στο τετράγωνο
cosine	Μέσος	Μεγιστοποίηση του αθροίσματος του συνημιτόνου της ομοιότητας ενός αντικειμένου από το κέντρο της συστάδας
Bregman divergence	Μέσος	Ελαχιστοποίηση του αθροίσματος της απόκλισης Bregman από ένα αντικείμενο με το κέντρο της συστάδας του

Παρατηρούμε ότι για την απόσταση Manhattan (L_1) και τον στόχο της ελαχιστοποίησης του αθροίσματος των αποστάσεων, το κατάλληλο κέντρο είναι ο διάμεσος των σημείων σε μια συστάδα. Η τελευταία εγγραφή στον πίνακα 3.2, η απόκλιση Bregman, είναι στην πραγματικότητα μια κλάση από μέτρα εγγύτητας που εμπεριέχει την τετραγωνική Ευκλείδεια απόσταση $(L_2)^2$, την απόσταση Mahalanobis, και την ομοιότητα συνημίτονου. Η σημασία των συναρτήσεων της απόκλισης Bregman είναι ότι κάθε μια συνάρτηση μπορεί να χρησιμοποιηθεί ως την βάση από έναν αλγόριθμο συσταδοποίησης τύπου k-means, με τον μέσο ως κέντρο. Συγκεκριμένα, αν χρησιμοποιήσουμε μια απόκλιση Bregman ως συνάρτηση εγγύτητας, τότε ο αλγόριθμος συσταδοποίησης θα έχει τις συνήθεις ιδιότητες ενός αλγορίθμου k-means σε σχέση με την απόκλιση, τοπικά ελάχιστα κτλ. Έτσι οι ιδιότητες ενός τέτοιου αλγορίθμου μπορεί να αναπτυχθεί για όλες τις

πιθανές αποκλίσεις Bregman. Στην πραγματικότητα οι αλγόριθμοι k-means που χρησιμοποιούν την ομοιότητα συνημίτονου ή την τετραγωνική Ευκλείδια απόσταση είναι συγκεκριμένες περιπτώσεις ενός γενικού αλγορίθμου συσταδοποίησης βασισμένου στις αποκλίσεις Bregman.

Όταν χρησιμοποιείται η τυχαία εκκίνηση των κέντρων, διαφορετικές εκτελέσεις του k-means παράγουν διαφορετικά SSE. Αυτό απεικονίζεται με τα σύνολα από δυσδιάστατα σημεία που έχει 3 φυσικές συστάδες από σημεία. Η εικόνα 3.11 (α) δείχνει μια λύση συσταδοποίησης που είναι το γενικό ελάχιστο του SSE για τρεις συστάδες ενώ στην (β) φαίνεται μια μη σωστή συσταδοποίηση που έχει μόνο τοπικό ελάχιστο.



Εικόνα 3.10: Βέλτιστη (α) και μη βέλτιστη (β) συσταδοποίηση
Πηγή: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Η επιλογή των κατάλληλων αρχικών κέντρων είναι το κλειδί της βασικής διαδικασίας του k-means. Μια κοινή προσέγγιση είναι να επιλεγούν τα αρχικά κέντρα τυχαία αν και τα αποτελέσματα της συσταδοποίησης είναι συχνά ανεπαρκή [16].

3.3.3 Πλεονεκτήματα-Μειονεκτήματα

Τα δυο βασικά πλεονεκτήματα του k-means είναι ότι είναι απλός και γρήγορος. Οι απαιτήσεις χωρητικότητας για τον k-means είναι μικρές γιατί αποθηκεύονται μόνο τα σημεία δεδομένων και τα κέντρα. Συγκεκριμένα ο χώρος που χρειάζεται είναι O((m + K)n), όπου m είναι ο αριθμός των σημείων και n είναι ο αριθμός των χαρακτηριστικών. Ο χρόνος που θα χρειαζόταν είναι O(I * K * m * n), όπου I είναι ο αριθμός των επαναλήψεων που χρειάζεται για την σύγκλιση.

Όπως έχει ήδη αναφερθεί το I είναι συνήθως μικρό και μπορεί να οριοθετηθεί συνήθως με ασφάλεια, καθώς οι περισσότερες αλλαγές συμβαίνουν στις πρώτες επαναλήψεις. Έτσι ο k-means είναι γραμμικός στο m ,τον αριθμό δηλαδή των σημείων και είναι αποτελεσματικός όπως απλός δεδομένου ότι το K ,ο αριθμός των συστάδων είναι σημαντικά μικρότερος από των αριθμό των m.

Στα μειονεκτήματα της χρήσης του k-means είναι ότι δεν μπορεί να διαχειριστεί μη σφαιρικές συστάδες ή με διαφορετικά μεγέθη και πυκνότητες, αν και μπορεί να βρει υποσυστάδες αν έχει οριστεί μεγάλος αριθμός συστάδων. Ο k-means έχει επίσης πρόβλημα στα δεδομένα που έχουν ακραίες τιμές (outliers). Τέλος, ο k-means περιορίζεται σε δεδομένα που υπάρχει η έννοια του κέντρου. Μια σχετική τεχνική η K-medoids δεν έχει αυτό τον περιορισμό αλλά χάνει σε κατανάλωση πόρων[16].

3.3.4 Ιδιαίτερα Ζητήματα

Ο k-means έχει κάποιες ιδιαιτερότητες που πρέπει να αποσαφηνιστούν πριν να προχωρήσουμε στην εφαρμογή του. Δεν μπορεί να διαχειριστεί **missing values**, δηλαδή στοιχεία τα οποία έχουν κενές τιμές σε κάποιες μεταβλητές. Αυτό αντιμετωπίζεται είτε συμπληρώνοντας όπου υπάρχουν κενά με τον μέσο όρο (αν είναι ποσοτική η μεταβλητή), είτε με την τίμη που συναντάται πιο συχνά (όταν είναι ποιοτική). Επίσης υπάρχουν οι επιλογές στα διάφορα προγράμματα να μπει το ελάχιστο ή το μέγιστο ή ακόμα και το μηδέν. Όλα αυτά επιλέγονται από τον χρήστη ανάλογα με τις ανάγκες.

Μια άλλη περίπτωση είναι όταν έχουμε ποσοτικές μεταβλητές με μεγάλη απόκλιση μεταξύ τους πχ ηλικία (1-100) και εισόδημα (μπορεί να φτάσει θεωρητικά ως το άπειρο). Για να έχουμε σε μια τέτοια περίπτωση πιο καλά αποτελέσματα προτείνεται να γίνει **κανονικοποίηση** (Normalization). Μια τιμή που της έχει γίνει κανονικοποίηση είναι μια τιμή που έχει επεξεργαστεί με έναν τρόπο που κάνει εφικτή την αποδοτική σύγκριση με άλλες τιμές. Ένα παράδειγμα κανονικοποίησης είναι η μετατροπή όλων το χαρακτήρων σε μικρά (ή κεφαλαία) για να αποκλειστούν οι διαφορές μεταξύ κεφαλαίων και μικρών που δεν μας ενδιαφέρουν. Με λίγα λόγια μετασχηματίζονται τα δεδομένα έτσι ώστε να μπορεί να γίνει απευθείας σύγκριση μεταξύ των μεταβλητών.

Ένα ακόμα ζήτημα είναι το πως γίνεται η επιλογή αρχικών σημείων. Ξεκινώντας από διαφορετικά σημεία μπορεί να έχει διαφορετικά αποτελέσματα στις τελικές συστάδες, οπότε πρέπει να δούμε πως καθορίζεται αυτό από πρόγραμμα σε πρόγραμμα.

ΚΕΦΑΛΑΙΟ 4
Μελέτη Περίπτωσης

4.1 Περιγραφή της επιχειρηματικής περίπτωσης

Αυτό το κεφάλαιο παρουσιάζει ένα βασικό σενάριο σε σχέση με την δημιουργία του προφίλ των πελατών μιας τράπεζας. Αυτό το σενάριο εργασίας βασίζεται στην χρήση των τεχνικών βαθμολόγησης και τα αποτελέσματά τους.

Το να γίνει αντιληπτή η συμπεριφορά του πελάτη και να μπορούν να προβλεφθούν οι προτιμήσεις και οι ενέργειές τους είναι πολύ σημαντικό για κάθε επιχείρηση. Συγκεκριμένα, αυτό ισχύει απολύτως σε έναν κλάδο που έχει τόσο υψηλή ανταγωνιστικότητα όπως είναι ο τραπεζικός κλάδος. Το πλεονέκτημα του να καλύπτεις τις ανάγκες των πελατών μπορεί να κάνει πολύ μεγάλη διαφορά μακροπρόθεσμα.

Η τμηματοποίηση των πελατών είναι μια πολύ χρήσιμη τεχνική για να καθοδηγείται η αλληλεπίδραση με τον πελάτη. Παραδοσιακά, η τμηματοποίηση γίνεται από ειδικούς της αγοράς που με κάποιον τρόπο δημιουργούν και περιγράφουν τομείς όπως θεωρούν κατάλληλα. Με την εξόρυξη πληροφορίας όμως, η τράπεζα μπορεί να ανιχνεύσει διαφορετικά μοτίβα στην συμπεριφορά των πελατών, χωρίς να βασίζεται στο ένστικτο. Ακόμα κι αν οι τεχνικές της εξόρυξης πληροφορίας χρησιμοποιηθούν για να βελτιώσουν τα κανάλια προς έναν μοναδικό τομέα, η τράπεζα ακόμα αντιμετωπίζει το ζήτημα του πώς να αξιοποιήσει καλύτερα τους υπαλλήλους γραφείου, στο να χρησιμοποιούν τα αποτελέσματα της εξόρυξης σε καθημερινές τους λειτουργίες. Εάν η εξόρυξη της πληροφορίας τεθεί σε εφαρμογή στις αλληλεπιδράσεις με τον πελάτη, ένα ίδρυμα μπορεί να επιτύχει την πλήρη επιστροφή χρημάτων του κόστους της επένδυσης πάνω στο data mining.

Για παράδειγμα τίθεται το πρόβλημα πώς να ξεχωρίσουμε τους πελάτες που αποφέρουν κέρδος στην τράπεζα από αυτούς που δεν το κάνουν. Πιο συγκεκριμένα, παρουσιάζουμε πως οι τελικοί χρήστες μπορούν εύκολα να έχουν πρόσβαση στις βαθμολογίες-scores των πελατών, μέσα από τις συνήθεις εφαρμογές. Επιλύουμε επίσης την προϋπόθεση ότι οι τελικοί χρήστες πρέπει να

μπορούν να παίρνουν τα σωστά αποτελέσματα που υπολογίστηκαν από τα πιο πρόσφατα δεδομένα σε σχέση με έναν συγκεκριμένο πελάτη. Αυτή η προσέγγιση επιτρέπει την βαθμολόγηση νέων πελατών.

Το άλλο σημαντικό ζήτημα εδώ, είναι ότι θέλουμε να βρίσκουμε τα προφίλ σε γρήγορους χρόνους. Η βαθμολόγηση στα υπάρχοντα προφίλ πρέπει να μπορεί να γίνεται πολύ γρήγορα και πολύ συχνά. Οι νέοι πελάτες συχνά επιλέγουν μια τράπεζα ή αλλάζουν ακόμα τράπεζες για προσωπικούς λόγους ή για λόγους δουλειάς. Συνεπώς, οι πιθανότητες είναι ότι μετά από μια επίσκεψη ,ή από δυο ενημερωτικές επισκέψεις, δεν θα επιστρέψουν ξανά αν οι υπηρεσίες ή τα προϊόντα δεν είναι στοχευμένα ή αναγνωρίσιμα σε αυτούς σε μια τιμή που μπορούν να αποδεχτούν. Στο συγκεκριμένο παράδειγμα, τίθεται το ερώτημα του πως θα γίνει το προφίλ και η βαθμολόγηση των νεοεισερχόμενων στην τράπεζα ,μέσω οποιουδήποτε καναλιού, με σχεδόν πραγματικού χρόνου βαθμολόγηση.[21]

4.2 Σκοπός / στόχοι μελέτης

Τα χαρακτηριστικά των υπηρεσιών της τράπεζας καθώς και οι τύποι των προϊόντων μπορούν να οδηγήσουν τους αναλυτές να επιλέξουν συγκεκριμένα τμήματα πελατών και να αγνοήσουν κάποια άλλα. Το προϊόν μιας υποθήκης μπορεί να είναι ενδιαφέρον μόνο σε ένα ή δυο τμήματα πελατών και όχι σε όλα. Τα δημογραφικά χαρακτηριστικά, όπως ο αριθμός μελών της οικογένειας, το οικογενειακό εισόδημα και η κατοχή ακίνητης περιουσίας μπορεί να οδηγήσει τον αναλυτή να επιλέξει ένα συγκεκριμένο τμήμα ως στοχευμένο.

Το χρονικό διάστημα που οι χρήστες είναι πελάτες μπορεί να οδηγήσει τον μάνατζερ της τράπεζας να αποφασίσει να επικεντρωθεί σε ένα υποσύνολο πελατών σε ένα τμήμα διαφορετικό από άλλους που ανήκουν στο ίδιο.

Το πρώτο στάδιο είναι να γνωρίσεις τους πελάτες καλύτερα:

- Πρέπει να υπάρχει μια καλύτερη γνώση των νέων πελατών μέσω αξιολόγησης-βαθμολόγησης(score).
- Να αξιολογούνται οι νέοι πελάτες στην βάση των χαρακτηριστικών τους και το κατά πόσο ταιριάζουν με το προφίλ συγκεκριμένων τμημάτων.
- Να μπορούν οι πελάτες να αλληλεπιδρούν με την τράπεζα μέσω διαφορετικών καναλιών (ATM, Internet, τηλεφωνικά κτλ).

- Η τράπεζα να μπορεί να προσφέρει διαφοροποιημένες υπηρεσίες και προϊόντα που με την πάροδο του χρόνου μπορεί να χρειάζονται προσαρμογές και επιπλέον χαρακτηριστικά.
- Να βοηθά τον αναλυτή της τραπέζης να γνωρίζει με ποιούς τύπους πελατών έχει να κάνει η τράπεζα απ την αρχή.
- Να λειτουργεί με πελατοκεντρικό τρόπο αντί με έμφαση στην υπηρεσία ή το προϊόν.

Τα μοντέλα τμημάτων παράγονται από τον αναλυτή που είναι υπεύθυνος για την εξόρυξη πληροφορίας μέσω ενός κατάλληλου λογισμικού, όπως είναι αυτά που συγκρίνονται στην παρούσα εργασία. Οι βαθμολογήσεις μπορούν να χρησιμοποιηθούν από τους υπόλοιπους υπαλλήλους. Μπορούν να ενσωματωθούν στις καθημερινές αναφορές για να αναγνωρίζονται οι διάφοροι τύποι πελατών και τα πιθανά ενδιαφέροντα τμήματα της αγοράς [21].

4.3 Περιγραφή μεταβλητών

Στον παρακάτω πίνακα 4.1 παρατίθενται οι μεταβλητές και τα χαρακτηριστικά τους.

Πίνακας 4.1

Πηγή: http://www.redbooks.ibm.com/redbooks/pdfs/sg246879.pdf

Αριθμός μεταβλητής	**Περιγραφή πεδίου**	**Όνομα πεδίου**	**Είδος, μονάδα μέτρησης**
1	Κωδικός πελάτη	Client_id	Ποσοτική
2	Ηλικία	Age	Ποσοτική, έτη
3	Φύλο	Sex	Ποιοτική
4	Πολιτεία που κατοικεί ο πελάτης	State	Ποιοτική
5	Μηνιαίο εισόδημα	Salary	Ποσοτική, δολάρια
6	Επάγγελμα	Profession	Ποσοτική(κάθε επάγγελμα κι ένας κωδικός)
7	Πλήθος ιδιοκτησίας σπιτιών	House_ownership	Ποσοτική
8	Πλήθος αυτοκινήτων	Car_ownership	Ποσοτική

9	Οικογενειακή κατάσταση	Marital_status	Ποιοτική
10	Αριθμός εξαρτώμενων	N_of_dependents	Ποσοτική
11	Αν έχει παιδιά	Has_children	Ποιοτική
12	Χρόνος ως πελάτης	Time_as_customer	Ποσοτική, χρόνια
13	Μηνιαίο ποσό checking	Checking_amount	Ποσοτική, δολάρια
14	Συνολικό ποσό από αυτόματες πληρωμές	Total_amount_automatic _payments	Ποσοτική, δολάρια
15	Μηνιαίο ποσό καταθέσεων	Savings	Ποσοτική, δολάρια
16	Μηνιαία αξία τραπεζικών κονδυλίων	Bank_funds	Ποσοτική, δολάρια
17	Μηνιαία αξία μετοχών	Stocks	Ποσοτική, δολάρια
18	Μηνιαία αξία κρατικών ομολόγων	Govern_bonds	Ποσοτική, δολάρια
19	Ποσό κέρδους ή ζημίας κονδυλιών (τραπεζικές καταθέσεις, μετοχές και ομόλογα)	Gain_loss_funds	Ποσοτική, δολάρια
20	Αριθμός υποθηκών	N_mortgages	Ποσοτική
21	Μηνιαίο ποσό υποθηκών	Mortgage amount	Ποσοτική
22	Μηνιαίο ποσό αναλήψεων	Money_monthly_overdrawn	Ποσοτική, δολάρια
23	Μέση αξία χρεωστικών συναλλαγών	AVG_debit_transact	Ποσοτική, δολάρια
24	Αριθμός επιταγών που γράφτηκαν τον μήνα	Monthly_checks_written	Ποσοτική
25	Αριθμός αυτόματων πληρωμών	N_automatic_payments	Ποσοτική
26	Αριθμός συναλλαγών μέσω υπαλλήλου	N_trans_teller	Ποσοτική

27	Αριθμός συναλλαγών μέσω ATM	N_trans_ATM	Ποσοτική
28	Αριθμός συναλλαγών μέσω Web bank	N_trans_Web_bank	Ποσοτική
29	Αριθμός συναλλαγών μέσω kiosk	N_trans_kiosk	Ποσοτική
30	Αριθμός συναλλαγών μέσω τηλ κέντρου	N_trans_Call_center	Ποσοτική
31	Μηνιαίο όριο πιστωτικής κάρτας	Credit_card_limits	Ποσοτική
32	Αν έχει ασφάλεια ζωής	Insurance	Ποιοτική

[21]

4.4 Επιλογή μεθόδου Εξόρυξης Πληροφορίας

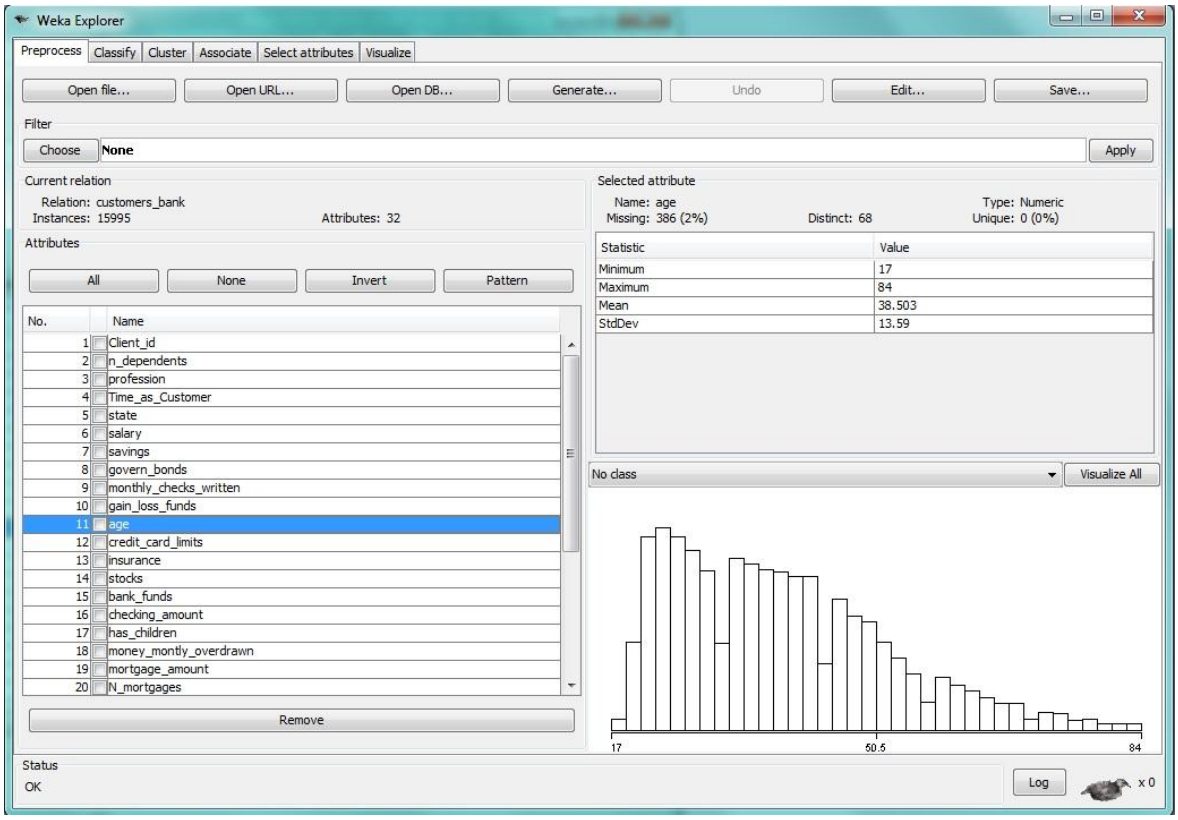
Ως μέθοδο εξόρυξης πληροφορίας θα χρησιμοποιήσουμε την συσταδοποίηση έτσι ώστε να δούμε τα διαφορετικά προφίλ πελατών για να μπορέσει η τράπεζα να ελέγξει κατά πόσο τα προϊόντα της έχουν το κατάλληλο κοινό ενδιαφέροντος και κατά πόσο μπορεί να μετατρέψει ή να δημιουργήσει κάποια προϊόντα έτσι ώστε να γίνουν πιο ελκυστικά στον πελάτη με βάση τις ανάγκες του. Η τράπεζα θα μπορεί να εφαρμόζει την συσταδοποίηση σε τακτά χρονικά διαστήματα, έτσι ώστε να παρατηρεί τυχόν αλλαγές που επιφέρει η εισαγωγή νέων πελατών. Δεδομένου ότι η τράπεζα έχει όλα τα στοιχεία που χρειάζεται για τον κάθε πελάτη στην βάση δεδομένων της, και επειδή η τράπεζα θέλει να ομαδοποιήσει τους πελάτες της σε σχέση με το προφίλ τους, η συσταδοποίηση επιλέγεται ως η καταλληλότερη τεχνική για την περίπτωση αυτή.

ΚΕΦΑΛΑΙΟ 5
Εφαρμογή Συσταδοποίησης

5.1 WEKA

5.1.1 Επιλογή μεταβλητών

Ξεκινώντας με το Weka επιλέγουμε το περιβάλλον **Explorer** και στην καρτέλα Preprocess επιλέγουμε το αρχείο .csv που υπάρχουν οι πελάτες με τα χαρακτηριστικά τους. Οι κυριότεροι τύποι αρχείων που χρησιμοποιούνται στο Weka είναι τα .arff .csv και .data .Υπάρχουν 32 μεταβλητές και 15995 εγγραφές στο αρχείο και μπορούμε να δούμε επιλέγοντας την κάθε μια μεταβλητή αν είναι numeric ή nominal. Στις περιπτώσεις ποσοτικών (numeric) μεταβλητών μας παρέχεται το ελάχιστο, το μέγιστο, ο μέσος όρος και η τυπική απόκλιση. Επίσης για κάθε μεταβλητή, μας δίνεται η πληροφορία για το πόσες τιμές υπάρχουν στο αρχείο, πόσες είναι μοναδικές και αν υπάρχουν εγγραφές που να λείπει η πληροφορία (missing values). Στις περιπτώσεις ποιοτικών (nominal) μεταβλητών μας εμφανίζει πόσες ετικέτες-επιλογές υπάρχουν και πόσες εγγραφές έχουν την κάθε επιλογή. Και στις ποσοτικές και στις ποιοτικές υπάρχει ένα γράφημα που μας δείχνει την κατανομή συχνοτήτων των εγγραφών για κάθε μεταβλητή. Στην Εικόνα 5.1 φαίνεται για την μεταβλητή age η κατανομή καθώς και το πώς είναι το περιβάλλον στην καρτέλα preprocess.



Εικόνα 5.1: Η καρτέλα Preprocess στο Περιβάλλον Explorer του Weka

Πριν προχωρήσουμε στην συσταδοποίηση, θα πρέπει να επιλέξουμε τις μεταβλητές που θα λάβουν μέρος στην εκτέλεση του αλγόριθμου, δηλαδή εκείνες τις μεταβλητές που έχουν πρακτική σημασία στην δημιουργία των προσδοκώμενων προφίλ. Το client_id πού είναι ο κωδικός του κάθε πελάτη δεν μας χρειάζεται στην συγκεκριμένη περίπτωση καθώς δεν έχει καμία αξία στην συσταδοποίηση για να το συμπεριλάβουμε. Επίσης, το profession στο συγκεκριμένο αρχείο είναι numeric. Το κάθε νούμερο αντιστοιχεί σε ένα επάγγελμα, όμως αυτή η αντιστοίχηση δεν είναι διαθέσιμη. Το να ερμηνεύσουμε σε ένα προφίλ μιας συστάδας ότι έχει μέσο όρο επαγγέλματος δεν έχει επίσης καμία αξία οπότε καταλήγουμε ότι είναι καλύτερο και αυτή η μεταβλητή να μην ληφθεί υπόψη. Οι υπόλοιπες μεταβλητές θα χρησιμοποιηθούν κανονικά για την συσταδοποίηση και είναι 30 στο σύνολο.

5.1.2 Μετασχηματισμός μεταβλητών

Το πρόγραμμα δίνει την δυνατότητα να χρησιμοποιηθούν και φίλτρα στα δεδομένα πριν την οποιαδήποτε χρήση τους. Υπάρχουν τα φίλτρα που

χρησιμοποιούνται στις μεταβλητές και φίλτρα που χρησιμοποιούνται στις εγγραφές.

Στα φίλτρα των μεταβλητών δίνεται η δυνατότητα να ενώσεις δυο τιμές, να μετατρέψεις numeric σε nominal ή binary, να αντιγράψεις μεταβλητές, να κανονικοποιήσεις όλες τις numeric μεταβλητές, να συμπληρώσεις τα missing values σε κάποιες μεταβλητές τους με τους αντίστοιχους μέσους όρους για τις ποσοτικές μεταβλητές και με τις αντίστοιχες επικρατούσες τιμές για τις ποιοτικές μεταβλητές. Η αναλυτική περιγραφή για το κάθε φίλτρο μεταβλητών παρέχεται στο http://weka.sourceforge.net/doc/weka/filters/unsupervised/attribute/package-summary.html .

Στα φίλτρα των εγγραφών δίνεται η δυνατότητα να κανονικοποιηθούν οι τιμές των εγγραφών, να επιλεγεί ένα τυχαίο δείγμα τους, Η αναλυτική περιγραφή για το κάθε φίλτρο εγγραφών παρέχεται στο:
http://weka.sourceforge.net/doc/weka/filters/unsupervised/instance/package-summary.html .

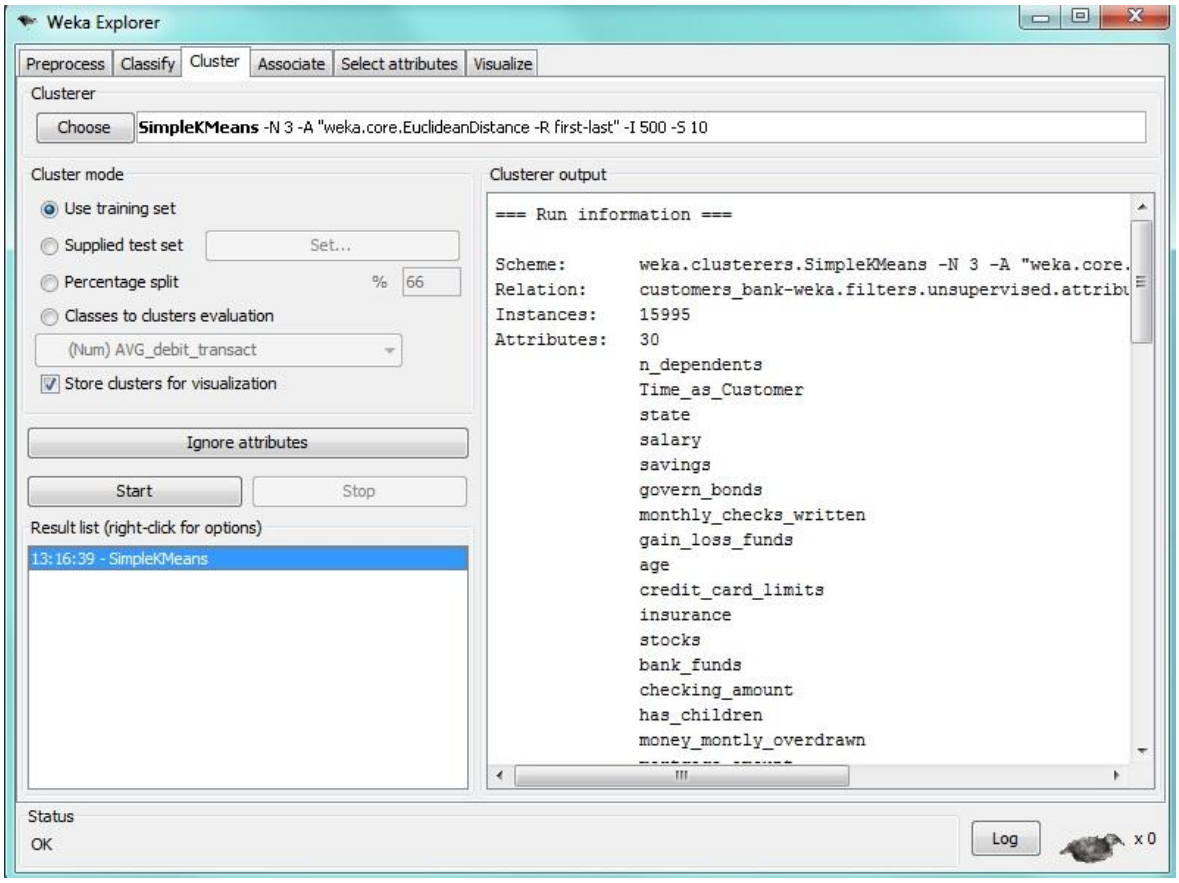
Στην περίπτωση των δεδομένων της παρούσας εργασίας δεν εφαρμόζεται κάποιος μετασχηματισμός καθώς τα δεδομένα είναι στην επιθυμητή μορφή. Εξαίρεση αποτελεί η συμπλήρωση των missing values διότι ο αλγόριθμος kmeans δεν μπορεί να διαχειριστεί missing values.

5.1.3 Προσδιορισμός παραμέτρων

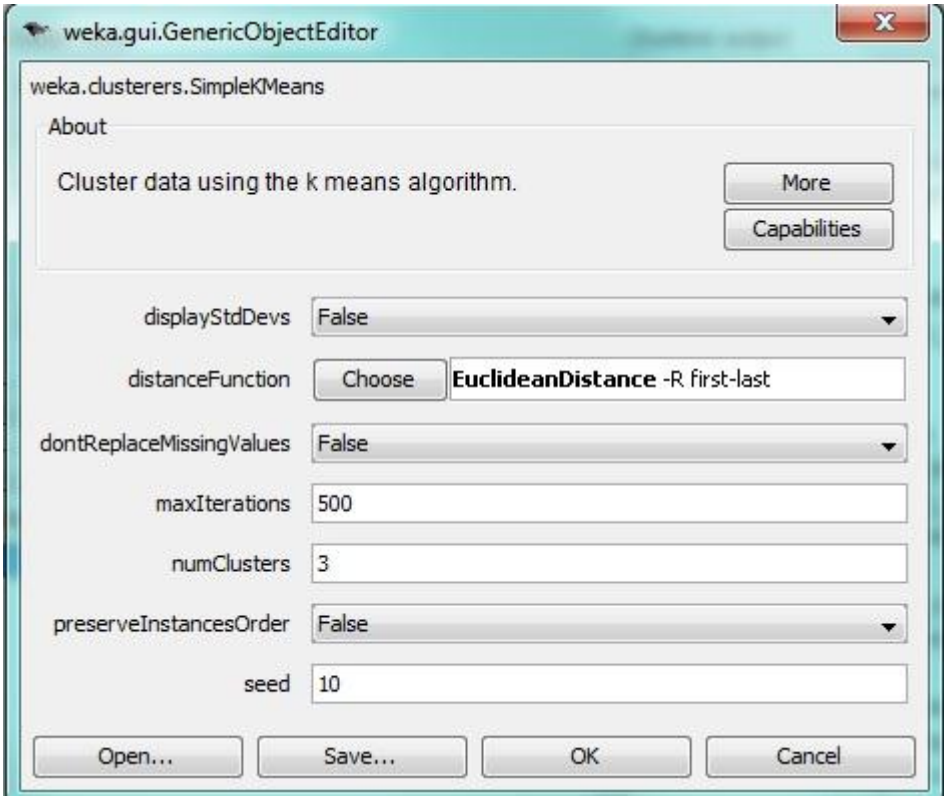
Έχοντας ολοκληρώσει τις ρυθμίσεις και την επιλογή των μεταβλητών επιλέγουμε την καρτέλα Cluster, στην οποία πραγματοποιείται η συσταδοποίηση. Στην εικόνα 5.2 φαίνεται το περιβάλλον αυτής της καρτέλας.

Μπορούμε να επιλέξουμε ανάμεσα σε 12 διαφορετικούς clusterers και εμείς θα χρησιμοποιήσουμε τον SimpleKMeans που είναι ο απλός kmeans. Έχοντας επιλέξει την επιλογή Store clusters for visualization θα μπορέσουμε αφού τρέξουμε τον αλγόριθμο να δούμε γραφήματα με τις συσχετίσεις ανάμεσα στις μεταβλητές.

Ας δούμε τώρα τις παραμέτρους που μπορούμε να ρυθμίσουμε για τον kmeans. Πατώντας πάνω στον clusterer μέσα στην γραμμή που εμφανίζεται το ποιος αλγόριθμος έχει επιλεγεί, μας ανοίγει ένα παράθυρο όπως φαίνεται στην εικόνα 5.3.



Εικόνα 5.2 : Η καρτέλα Cluster στο Περιβάλλον Explorer του Weka



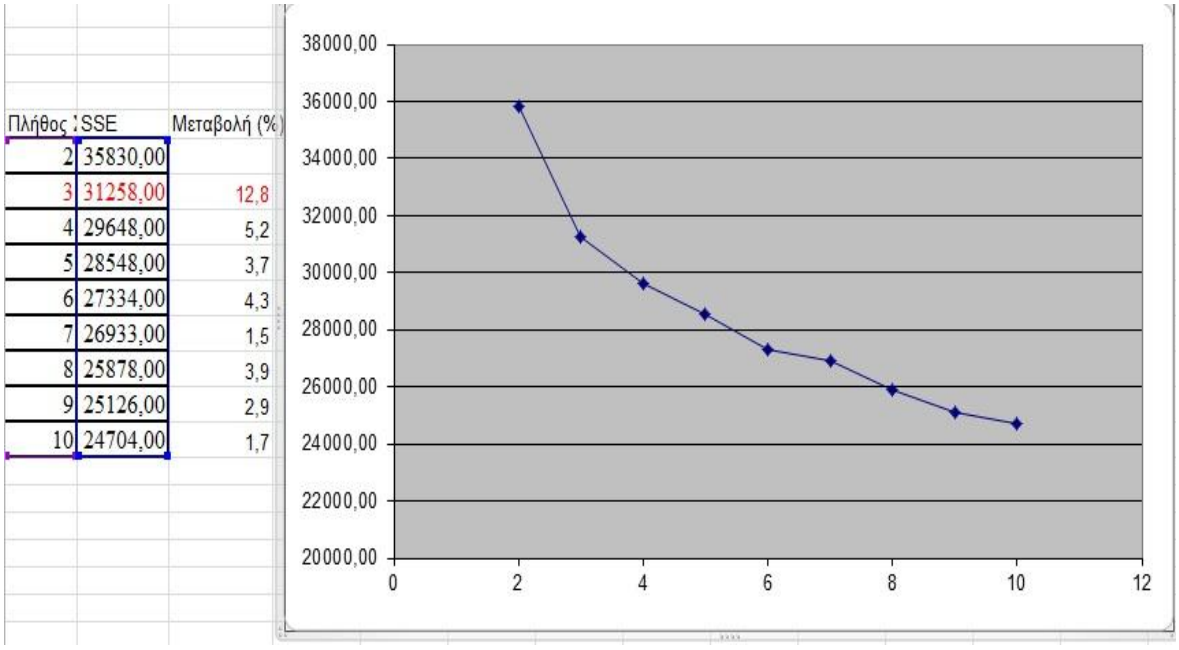
Εικόνα 5.3 : Η παραμετροποίηση του kmeans στο Περιβάλλον Explorer του Weka

- Στην επιλογή *displayStdDevs,* όταν έχουμε επιλέξει True τότε στα αποτελέσματα του αλγορίθμου μας εμφανίζει τις τυπικές αποκλίσεις των αριθμητικών μεταβλητών και το πόσες φορές εμφανίστηκε η κάθε μια επιλογή των nominal μεταβλητών.
- Στην επιλογή *distanceFunction* επιλέγουμε με ποιόν τρόπο θέλουμε να μετριέται η απόσταση-σύγκριση ανάμεσα στις εγγραφές. Υπάρχουν 4 επιλογές εκ των οποίων οι δυο πιο διαδεδομένες είναι η Ευκλείδεια και η Μανχάταν.
- Στην επιλογή *dontReplaceMissingValues* όταν είναι επιλεγμένο το False τότε αντικαθιστά όπου βρει κενά σε μεταβλητές με τον μέσο όρο αυτής της μεταβλητής.
- Στην επιλογή *maxIterations* ορίζουμε τον μέγιστο αριθμό επαναλήψεων.
- Στην επιλογή *numClusters* ορίζουμε τον αριθμό των clusters που θέλουμε να τρέξει ο αλγόριθμος.
- Στην επιλογή *preserveInstancesOrder* όταν είναι επιλεγμένο το True, τότε διατηρεί την σειρά των εγγραφών.
- Στην επιλογή *seed* ορίζουμε τον τυχαίο αριθμό seed που θα χρησιμοποιηθεί.

Επιλέγουμε False στην πρώτη επιλογή γιατί δεν μας ενδιαφέρει να εμφανιστεί η τυπική απόκλιση, επιλέγουμε την Ευκλείδεια απόσταση ως μέτρο, επιλέγουμε false στο dontReplaceMissingValues γιατί θέλουμε να τις αντικαταστήσει (ο kmeans δεν μπορεί να διαχειριστεί κενά στις εγγραφές), ως max στις επαναλήψεις αφήνουμε το default που είναι το 500, όπως αφήνουμε και το 10 στο seed. Δεν μας ενδιαφέρει στην συγκεκριμένη περίπτωση η σειρά των εγγραφών, οπότε το αφήνουμε false στο preserveInstancesOrder. Τέλος, ως προς το πόσες συστάδες θέλουμε να τρέξει ο αλγόριθμος δεν μπορούμε να το ξέρουμε εκ των προτέρων, οπότε πρέπει να το αναλύσουμε λίγο περισσότερο.

Όταν πραγματοποιούμε μία συσταδοποίηση στο Weka, στα αποτελέσματα που παράγονται συμπεριλαμβάνεται η τιμή του "*Within cluster sum of squared errors*". Στην εφαρμογή αυτή επιλέχθηκε να επαναληφθεί ο αλγόριθμός για διαφορετικό πλήθος συστάδων (numCluster: 2 έως 10) και να καταγραφεί η αντίστοιχη τιμή του SSE. Ο κανόνας για να επιλέξουμε το καταλληλότερο πλήθος συστάδων είναι να εντοπίσουμε εκείνο το k για το οποίο η μεταβολή στην τιμή του

SSE αρχίζει να είναι λίγότερο «απότομή». Όπως βλέπουμε στην εικόνα 5.4, η μεταβολή από τους 2 clusters στους 3 είναι η μεγαλύτερη πριν τελικά αρχίσει η μείωση να τείνει να είναι ηπιότερη σε κάθε αλλαγή του πλήθους των συστάδων. Συνεπώς, το πλήθος των συστάδων επιλέγεται να είναι ίσο με τρία (3).

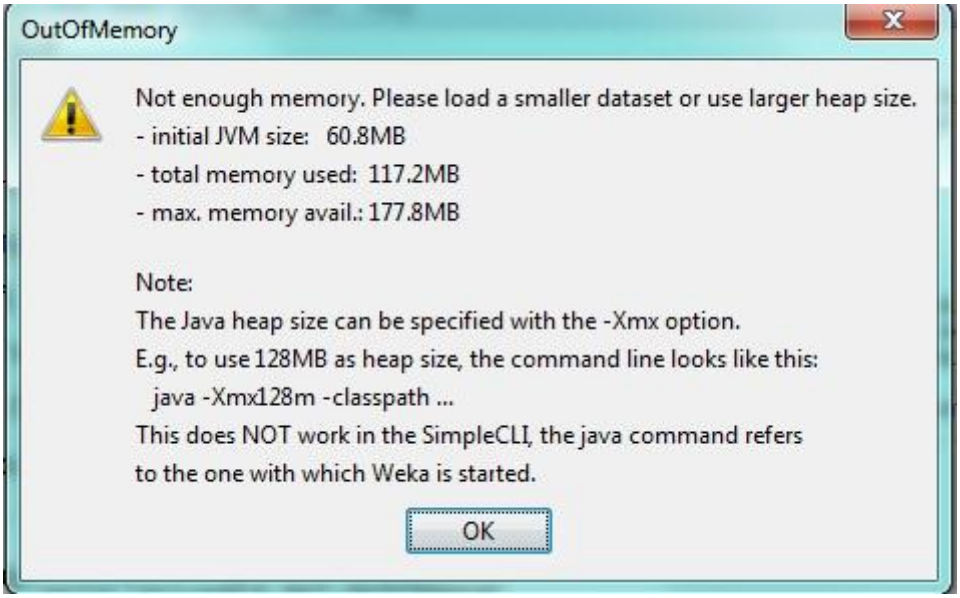


Πλήθος	SSE	Μεταβολή (%)
2	35830,00	
3	31258,00	12,8
4	29648,00	5,2
5	28548,00	3,7
6	27334,00	4,3
7	26933,00	1,5
8	25878,00	3,9
9	25126,00	2,9
10	24704,00	1,7

Εικόνα 5.4 : Το ποσοστό μεταβολής του SSE είναι μεγαλύτερο στις 3 συστάδες

5.1.4 Εφαρμογή και αποτελέσματα

Στην φάση αυτή, επιλέγουμε την εφαρμογή του T kmeans με 3 συστάδες και στο clusterer output παίρνουμε τα αποτελέσματα σε μορφή text. Στο σημείο αυτό, θα πρέπει να αναφερθεί ένα πρόβλημα που προέκυψε με το WEKA σε αυτή την φάση. Όταν ξεκινούσε να τρέξει την συσταδοποίηση εμφανιζόταν το μήνυμα **out of memory error** και στην συνέχεια το πρόγραμμα έκλεινε (Εικόνα 5.5).



Εικόνα 5.5 : Το OutOfMemory error

Επειδή ο Explorer του Weka τρέχει εξ ολοκλήρου στην μνήμη RAM, στην ουσία ενημέρωνε ότι δεν επαρκούσε η μνήμη. Δεδομένου ότι στον υπολογιστή που έτρεξε το πρόγραμμα η μνήμη RAM είναι 3GB ψάξαμε να βρούμε στα manuals και σε κάποιο επίσημο φόρουμ την λύση του προβλήματος. Βρήκαμε στο troubleshooting του επίσημου site του Weka ότι μπορούμε να επεκτείνουμε την Ram που δεσμεύει το πρόγραμμα σε περίπτωση που εμφανίσει αυτό το πρόβλημα και προτεινόταν να αλλάξουν οι by default επιλογές της Java στον υπολογιστή έτσι ώστε να δεσμεύεται περισσότερη RAM. Ψάχνοντας λίγο ακόμα στο διαδίκτυο κάποιος χρήστης του Weka σε forum εκτός του επίσημου πρότεινε να αλλαχτεί ο κώδικας του αρχείου RunWeka.ini ,που είναι το αρχείο των settings του προγράμματος, το maxheap= με τα mb της Ram που θα δεσμεύονται με την έναρξη του περιβάλλοντος Explorer στο πρόγραμμα. Ορίζοντάς το στα 1024mb και αποθηκεύοντας το αρχείο δεν ξαναεμφανίστηκε το παραπάνω error αλλά δέσμευε πλέον το πρόγραμμα 1Gb Ram κατά την εκτέλεσή του.

Αφού ορίσαμε λοιπόν τις παραμέτρους όπως είναι στην εικόνα 5.3 τρέξαμε τον kmeans με το κουμπί Start και πήραμε τα εξής αποτελέσματα:

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-
last" -I 500 -S 10
Relation:     customers_bank-weka.filters.unsupervised.attribute.Remove-R1,3
Instances:    15995
Attributes:   30
                n_dependents
                Time_as_Customer
                state
                salary
                savings
                govern_bonds
                monthly_checks_written
                gain_loss_funds
                age
                credit_card_limits
                insurance
                stocks
                bank_funds
                checking_amount
                has_children
                money_montly_overdrawn
               mortgage_amount
               N_mortgages
               N_trans_Call_center
               n_automatic_payments
               Total_ammount_automatic_payments
               N_trans_WEb_bank
               N_trans_kiosk
               N_trans_teller
               car_ownership
               house_ownership
               N_trans_ATM
               marital_status
               SEX
               AVG_debit_transact
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 12
```

Within cluster sum of squared errors: 31258.065045819967
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (15995)	Cluster# 0 (4724)	1 (5927)	2 (5344)
n_dependents	2.06	2.2475	1.8271	2.1527
Time_as_Customer	2.4511	2.4384	2.5323	2.3722
State	NY	MI	NY	NY
Salary	2043.4367	1968.8548	380.7729	3953.4167
Savings	30370.6513	14352.961	1912.6364	76092.6243
govern_bonds	5607.0945	3951.1461	2751.8249	10237.6864
monthly_checks_written	4.518	5.3605	3.4974	4.9051
gain_loss_funds	57.0334	50.5017	17.1822	107.0062
age	38.5032	37.6534	34.046	44.1978
credit_card_limits	1261.8631	1270.0042	1059.1362	1479.5097
insurance	0.7293	0.6914	0.8421	0.6377
stocks	3140.489	1772.8755	1283.0645	6409.4936
bank_funds	2626.8547	1885.6372	788.4869	5321.0009
checking_amount	1098.0124	1056.6215	611.0202	1674.7216
has_children	0.5	0.4865	0.2313	0.8099
money_montly_overdrawn	52.3821	42.5658	21.1014	95.7528
mortgage_amount	2047.6554	2008.6213	375.9275	3936.2648
N_mortgages	0.8086	1.0531	0.3685	1.0805
N_trans_Call_center	0.0317	0.0243	0.0091	0.0632
n_automatic_payments	0.0738	0.0599	0.0236	0.1418
Total_ammount_automatic _payments	4.4162	2.9469	0.8935	9.622
N_trans_WEb_bank	1.9404	2.1022	1.3521	2.4497
N_trans_kiosk	1.9231	1.8804	1.8649	2.0254
N_trans_teller	1.8084	1.9414	1.2109	2.3536
car_ownership	0.9324	1	0.8174	1
house_ownership	0.8088	1.0531	0.3685	1.081
N_trans_ATM	2.8794	3.0491	1.8448	3.8769
marital_status	MARRIED	MARRIED	SINGLE	DIVORCED
SEX	M	M	M	F
AVG_debit_transact	1360.4318	1472.2794	457.5934	2262.8933

Clustered Instances

0 4724 (30%)
1 5927 (37%)
2 5344 (33%)

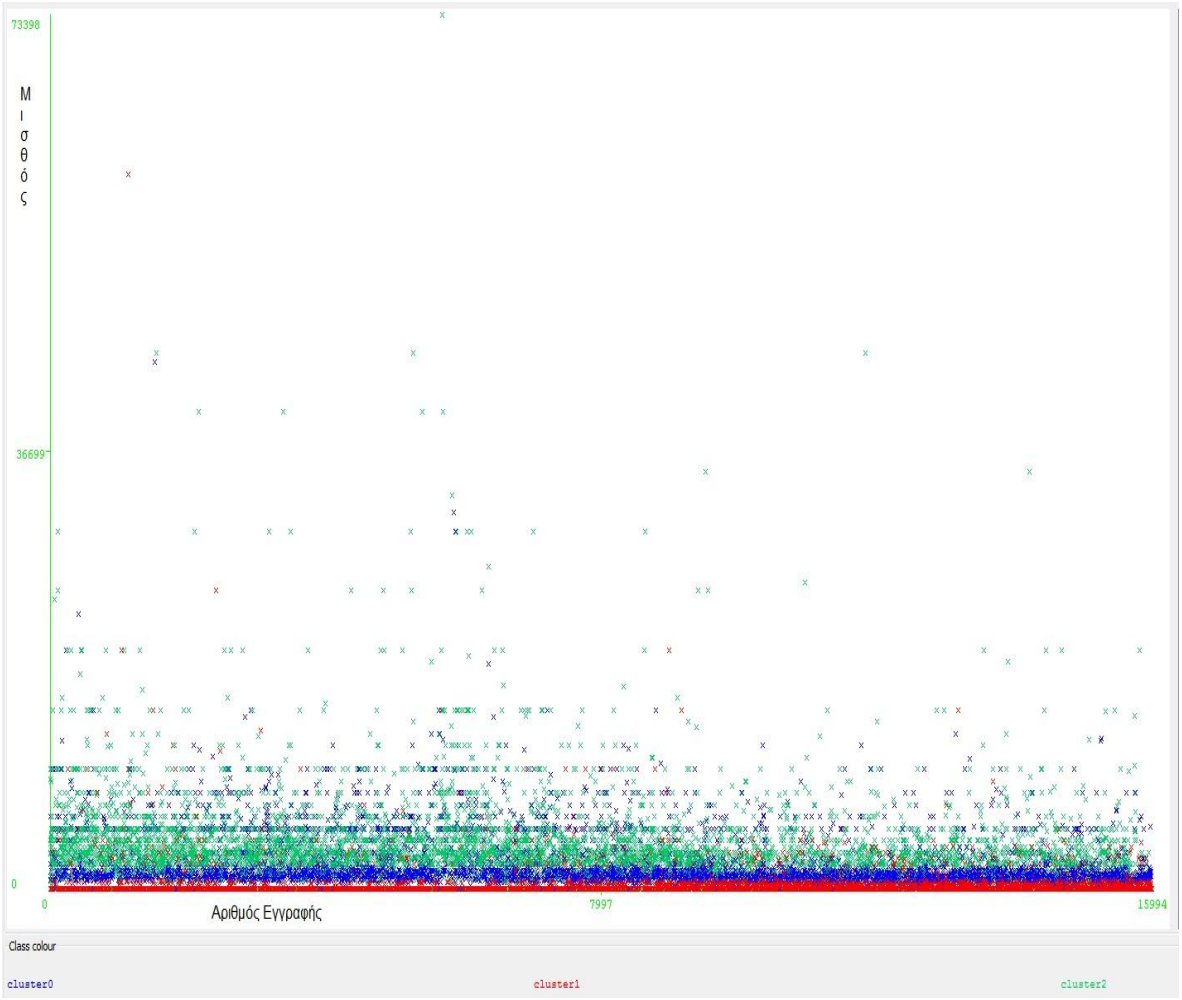
Όπως φαίνεται στα αποτελέσματα, έχουνε δημιουργηθεί 3 διαφορετικά προφίλ πελατών της τράπεζας (3 συστάδες). Με μια πρόχειρη ματιά βλέπουμε ότι τα προφίλ των τριών συστάδων διαφέρουν πάρα πολύ στον μισθό (Salary), στις καταθέσεις (Savings) και στο γενικό κέρδος (gain_loss_funds). Όλα αυτά μπορούμε να τα δούμε και σχηματικά κάνοντας δεξί κλικ στο Result list πάνω στο αποτέλεσμά μας και επιλέγοντας Visualize Cluster Assignments.

Ας δούμε τώρα το προφίλ της συστάδας 0 που περιέχει 4724 εγγραφές. Είναι άντρας παντρεμένος, έχει αυτοκίνητο, είναι 37 χρονών, κατέχει 1 σπίτι, έχει μια υποθήκη, το 48% έχει τουλάχιστον ένα παιδί, το 69% έχει ασφάλεια, ζει κατά βάση στην πολιτεία του Michigan, έχει αποταμιεύσεις 14352 δολάρια ,μισθό 1968 δολάρια, είναι πελάτης της τραπέζης περίπου 2μιση χρόνια και έχει μηνιαία αξία τραπεζικών κονδυλίων 1885 δολάρια.

Η συστάδα 1 περιέχει 5927 εγγραφές. Το προφίλ του πελάτη είναι ως εξής: Είναι άντρας ανύπαντρος, έχει κατά 81% αυτοκίνητο, είναι 34 χρονών, δεν κατέχει σπίτι (0,36 σπίτια μέσο όρο), ως εκ τούτου δεν έχει υποθήκη, μόνο το 23% έχει παιδί, το 84% έχει ασφάλεια, ζει στην πολιτεία της Νέας Υόρκης, έχει αποταμιεύσεις 1912 δολάρια, μισθό 380 δολάρια ,είναι πελάτης της τραπέζης 2μιση χρόνια και έχει μηνιαία αξία τραπεζικών κονδυλίων 788 δολάρια.

Η συστάδα 2 περιέχει 5344 εγγραφές. Το προφίλ του πελάτη είναι ως εξής: Είναι γυναίκα με διαζύγιο, έχει αυτοκίνητο, είναι 44 χρονών, κατέχει ένα σπίτι, έχει μια υποθήκη, το 81% έχει ένα τουλάχιστον παιδί, το 63% έχει ασφάλεια, ζει στην πολιτεία της Νέας Υόρκης, έχει αποταμιεύσεις 76092 δολάρια, μισθό 3953 δολάρια, είναι πελάτης της τραπέζης 2 χρόνια και 4 μήνες και έχει μηνιαία αξία τραπεζικών κονδυλίων 5321 δολάρια.

Όπως βλέπουμε και στα προφίλ η μεταβλητές που είναι οι πιο διαφοροποιημένες είναι ο μισθός και οι αποταμιεύσεις. Ας δούμε τώρα στην εικόνα 5.6 το σχήμα που μας παρέχει το Weka.



Εικόνα 5.6: Ο μισθός ανά εγγραφή και με χρωματισμένα σε ποια συστάδα ανήκουν

Όπως φαίνεται και στο σχήμα οι πολύ χαμηλοί μισθοί ανήκουν στον cluster1 τα μεσαία εισοδήματα στον cluster0 και τα υψηλότερα ανήκουν κυρίως στον cluster2. Αντίστοιχο είναι και το γράφημα με τις αποταμιεύσεις με τις μικρές στον 1 τις μεσαίες στον 0 και τις υψηλότερες στον 2.

Στο παράθυρο του Visualize Cluster Assignments δίνεται η δυνατότητα να οριστεί οποιαδήποτε από τις μεταβλητές ως άξονας X είτε ως Y. Επίσης, δίνεται η δυνατότητα να χρωματιστούν διαφορετικά, όχι μόνο τα στοιχεία των συστάδων αλλά και οποιαδήποτε από τις μεταβλητές λαμβάνοντας διαγραμματικά όλες τις πληροφορίες που χρειάζονται για την εξαγωγή χρήσιμων συμπερασμάτων.

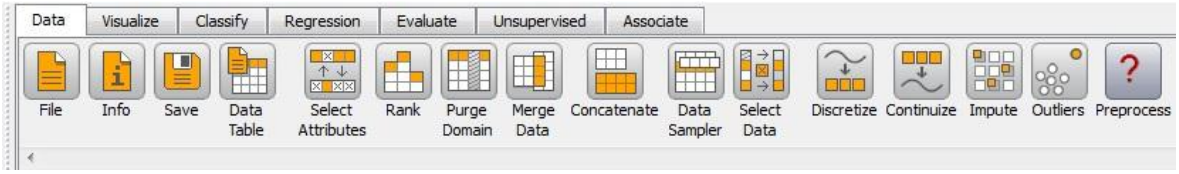
5.2 Orange

5.2.1 Επιλογή μεταβλητών

Ανοίγοντας το Orange, ανοίγει ένα παράθυρο στο οποίο δίνεται η δυνατότητα να προστεθούν διάφορα widgets, τα οποία μπορούν να ενωθούν έτσι ώστε να δημιουργηθεί η ροή ενός αλγορίθμου με σειριακό τρόπο. Στην περίπτωσή μας ξεκινάμε προσθέτοντας το widget "*File*". Κάνοντας δεξί κλικ πάνω στον canvas (έτσι έχουν ονομάσει οι δημιουργοί το παράθυρο που προσθέτεις τα widgets) από τις επιλογές Data επιλέγουμε το File και το widget προστίθεται στον canvas. Τώρα πρέπει να ορίσουμε το αρχείο που θέλουμε να εισαχθεί. Με διπλό κλικ πάνω στο File ανοίγει ένα παράθυρο και από εκεί επιλέγουμε μέσω της αναζήτησης το αρχείο customers_bank.csv.

Κλείνουμε το παράθυρο και επιλέγουμε πάνω στον κύκλο που υπάρχει δεξιά από το widget "*File*", επιλέγουμε Data και μετά "*Select Attributes",* το οποίο εισάγεται στον canvas έχοντας ως είσοδο το widget "*File*". Με διπλό κλικ πάνω στο "*Select Attributes*", ανοίγει ένα νέο παράθυρο το οποίο μας έχει σε μια στήλη όλα τις διαθέσιμες μεταβλητές και επιλέγουμε να περαστούν όλες εκτός από τις δυο Client_id και profession για τους λόγους που εξηγήθηκαν στο κεφάλαιο 5.1.1. Έχουμε έτσι ολοκληρώσει την εισαγωγή του αρχείου και την επιλογή των μεταβλητών που θέλουμε.

5.2.2 Μετασχηματισμός Μεταβλητών



Εικόνα 5.7: Τα διάφορα widgets που υπάρχουν για την διαχείριση των Data και των μεταβλητών

Ας αναλύσουμε λίγο τι δυνατότητες υπάρχουν στο πρόγραμμα σε σχέση με τον μετασχηματισμό των μεταβλητών αλλά και των δεδομένων. Στην εικόνα 5.7 βλέπουμε όλα τα widgets που έχουν να κάνουν με τα δεδομένα.

- Με το "*File*" εισάγεται το επιθυμητό αρχείο
- Το "*Info*" μπαίνει ως έξοδος του "*File*" και είναι ένα παράθυρο που ενημερώνει πόσες εγγραφές έχει, πόσες μεταβλητές κτλ.

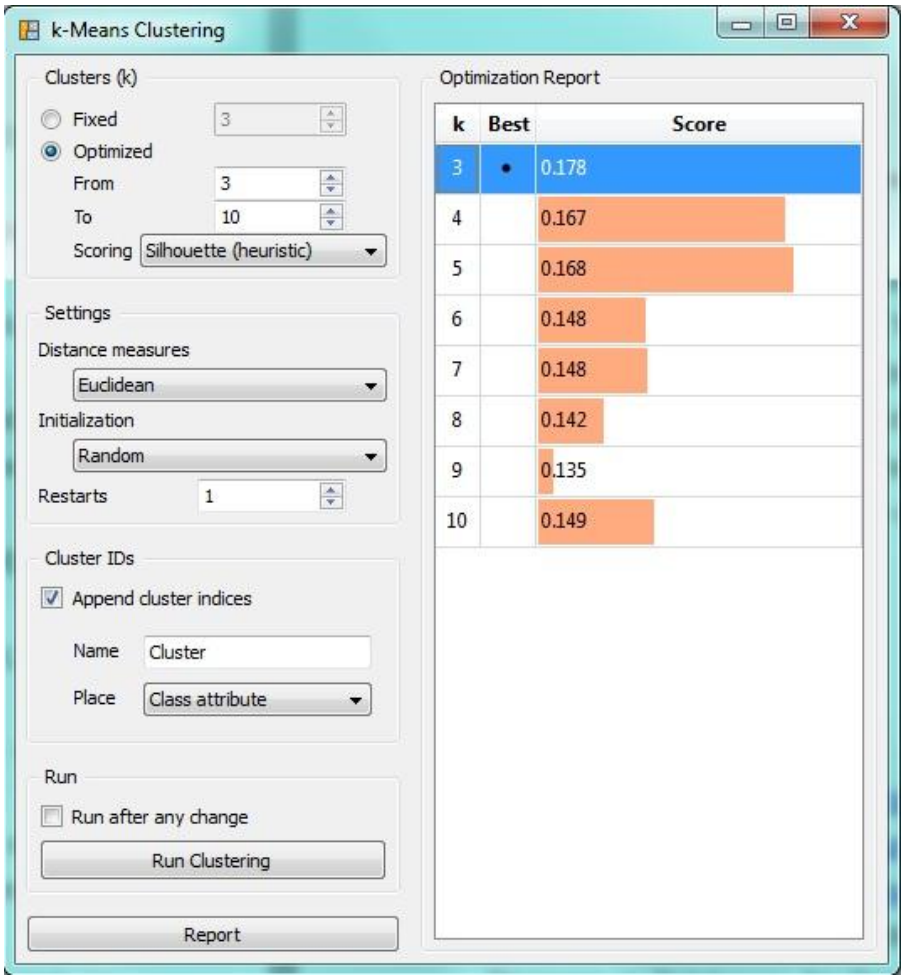
- Με το “*Save*” αποθηκεύονται τα δεδομένα σε νέο αρχείο ως .csv ή ως .tab
- Με το “*Data Table*” και ως είσοδο κάποιο αρχείο εμφανίζονται τα δεδομένα σε μορφή λογιστικού φύλλου (πχ Excel)
- Με το “*Select Attributes*” επιλέγονται οι μεταβλητές που χρειάζονται από το αρχείο που έχει ως είσοδο.
- Το “*Rank*” είναι ένα widget για να ταξινομηθούν οι μεταβλητές και να επιλεγεί ένα υποσύνολο μεταβλητών
- Με το “*Purge Domain*” δίνεται η δυνατότητα να αφαιρεθούν οι αχρησιμοποίητες τιμές των μεταβλητών / μεταβλητές και να ταξινομηθούν οι εναπομείνασες τιμές.
- Το “*Merge Data*” ενώνει 2 σύνολα δεδομένων βασισμένο στις τιμές των επιλεγμένων μεταβλητών
- Το “*Concatenate*” προσθέτει δεδομένα από πολλαπλές πηγές εισόδου
- Το “*Data Sampler*” επιλέγει ένα υποσύνολο εγγραφών από το εισαγόμενο data set
- Το “*Select Data*” επιλέγει εγγραφές από τον εισαγόμενο data set βασιζόμενο σε συνθήκες που έχει ορίσει ο χρήστης στις μεταβλητές (πχ όπου είναι μεγαλύτερο από 1000 δολάρια στο salary)
- Το “*Discretize*” «διαφοροποιεί» τις συνεχείς μεταβλητές από το εισαγόμενο data set
- Το “*Continuize*” μετατρέπει τις διακριτές μεταβλητές σε συνεχείς πλασματικά (αντίστροφη διαδικασία από το Discretize)
- Το “*Impute*” αντικαθιστά άγνωστες τιμές στα δεδομένα (είτε με τον μέσο όρο είτε με τυχαίες τιμές ή απλά ορίζεται να αγνοήσει τις εγγραφές με missing values)
- Με το “*Outliers*” γίνεται απλός εντοπισμός ακραίων τιμών συγκρίνοντας αποστάσεις μεταξύ παραδειγμάτων
- Με το “*Preprocess*” δίνεται η δυνατότητα να πραγματοποιηθεί μείξη όλων των υπολοίπων widgets που προαναφέρθηκαν.

Στην περίπτωση μας, δεν χρειάζεται να χρησιμοποιήσουμε κάποιο επιπλέον widget πέραν του “*Select Attributes*”. Εφόσον το επιθυμούμε, μπορούμε να

βάλουμε ως έξοδο στο “*File*” το widget “*Attribute Statistics*” από την καρτέλα Visualize για να δούμε τα στατιστικά στοιχεία των δεδομένων που έχουμε στο .csv αρχείο μας.

5.2.3 Προσδιορισμός παραμέτρων

Επιλέγουμε την έξοδο του “*Select Attributes*” και από το μενού Unsupervised την επιλογή “*k-means clustering*”. Με αυτόν τον τρόπο προστίθεται το widget με τον k-means. Παρατηρούμε ότι δεν υπάρχει ξεχωριστό μενού clustering και παρέχει 3 διαφορετικούς τύπους clustering: k-means, hierarchical και network. Ας δούμε τώρα τι παραμέτρους μπορούμε να δώσουμε στο πρόγραμμα για τον k-means. Με διπλό κλικ πάνω στο widget του k-means ανοίγει ένα παράθυρο όπως φαίνεται στην εικόνα 5.8.



Εικόνα 5.8: Το παράθυρο παραμετροποίησης του αλγορίθμου kmeans στο Orange

Καταρχήν μπορεί να οριστεί, όπως είναι φυσικό, πόσες συστάδες επιθυμούμε να δημιουργηθούν, αλλά παρέχει κι έναν δικό του «μηχανισμό», με 4 διαφορετικές μεθόδους, ο οποίος προτείνει το πλήθος των συστάδων. Για παράδειγμα, στην περίπτωση της εικόνας 5.8, επιλέχθηκε το πλήθος των συστάδων να είναι από 3 έως 10, και η μέθοδος να είναι η προεπιλεγμένη του λογισμικού (Silhouette – heuristic). Στο δεξιό μέρος της εικόνας, παρουσιάζεται η κατάταξη του πλήθους των συστάδων με βάση την αξιολόγησή τους από την αντίστοιχη μέθοδο. Όπως βλέπουμε, οι 3 συστάδες έχουν την υψηλότερη βαθμολογία.

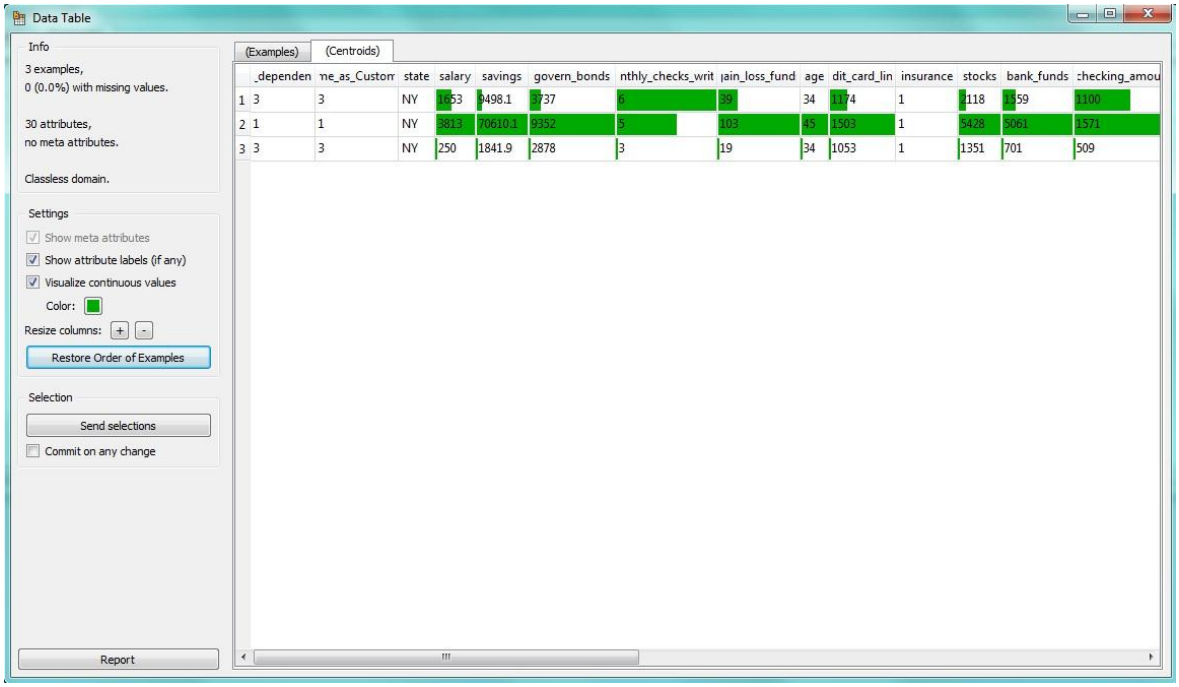
Στην συνέχεια, δίνεται η δυνατότητα επιλογής του μέτρου απόστασης που θα χρησιμοποιηθεί. Δίνονται 6 επιλογές με προεπιλεγμένη την Ευκλείδεια απόσταση (οι υπόλοιπες είναι: Pearson Correlation, Spearman Rank Correlation, Manhattan, Maximal και Hamming). Στο πεδίο Initialization υπάρχουν 3 επιλογές: Random(default), Diversity και Agglomerative clustering. Έχει ένα ακόμα πεδίο το Cluster Ids με το οποίο έχοντας επιλεγμένο το Append cluster indices δίνεται ένα όνομα και δημιουργείται μια νέα μεταβλητή Class στα αποτελέσματα έτσι ώστε να φαίνεται που ανήκει η κάθε εγγραφή μετά την συσταδοποίηση.

Ο αλγόριθμος εφαρμόζεται από την αρχή με κάθε αλλαγή εκτός αν απο-επιλέξουμε την επιλογή Run after any change. Ο αλγόριθμος τρέχει όταν πατάμε το κουμπί Run clustering αλλά το πώς παίρνουμε αποτελέσματα θα το δούμε στην αμέσως επόμενη ενότητα.

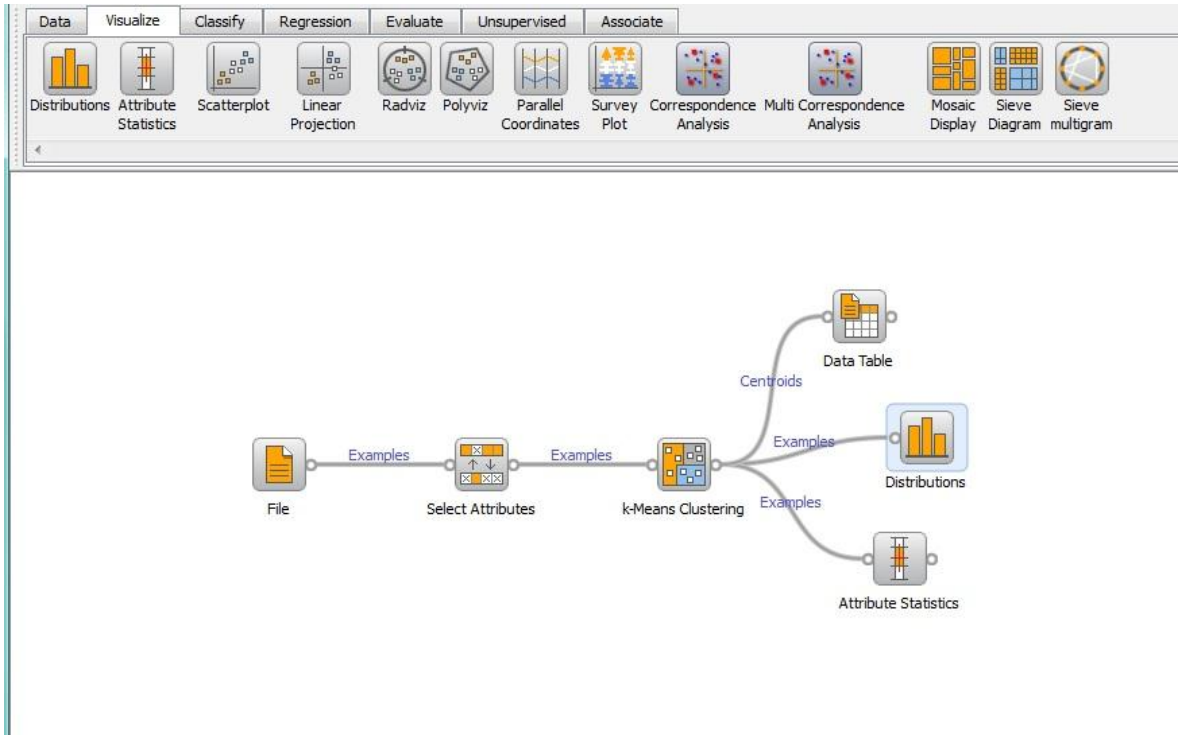
5.2.4 Εφαρμογή και αποτελέσματα

Ως έξοδο του “*k-means clustering*” ορίζουμε το “*Data Table*” από το μενού Data. Μας ανοίγει ένα παράθυρο που μας δίνει την δυνατότητα να επιλέξουμε να εμφανιστούν τα αποτελέσματα ως Examples, που θα μας εμφανίσει έτσι όλες τις εγγραφές σε μορφή πίνακα ή να επιλέξουμε το Centroids που θα μας βγάλει αποτελέσματα σε μορφή που τα θέλουμε στην προκειμένη περίπτωση. Δηλαδή, ανά συστάδα τι αποτελέσματα έχουμε για κάθε μεταβλητή. Τα αποτελέσματα που παίρνουμε διπλοπατώντας πάνω στο “*Data Table*” είναι σε μορφή πίνακα και το παράθυρο είναι όπως φαίνεται στην εικόνα 5.9. Αν θέλουμε να τα πάρουμε σε μορφή κειμένου πατάμε στο κουμπί κάτω αριστερά Report αλλά τα αποτελέσματα δεν είναι καθόλου ευανάγνωστα. Επίσης δεν είναι δυνατόν να τα κάνουμε copy paste ούτε ως text ούτε ως πίνακα.

Ως έξοδο στον “*k-means clustering*” βάζουμε επίσης και το “*Distributions*” από το μενού Visualize με επιλογή Examples στα αποτελέσματα (default) και το widget “*Attribute Statistics*” επίσης με επιλογή Examples (default). Έτσι έχουμε ολοκληρώσει την μορφή των widgets στο παράδειγμά μας. Στην εικόνα 5.10 φαίνεται πως είναι το κεντρικό παράθυρο ή αλλιώς Orange Canvas.

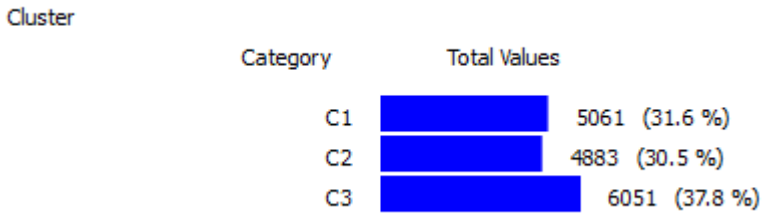


Εικόνα 5.9: Τα αποτελέσματα σε μορφή πίνακα με το widget “Data Table”



Εικόνα 5.10: Η τελική μορφή των συνδέσεων των widget στο παράδειγμά μας

Ας δούμε τώρα τις 3 συστάδες και τα προφίλ που δημιουργούνται. Από τα αποτελέσματα του "*Data Table*" παίρνουμε στρογγυλοποιημένα, χωρίς ακρίβεια κατ επέκταση, τους μέσους όρους από κάθε μια μεταβλητή. Οι διαφορές ανάμεσα στις συστάδες γίνονται εύκολα αντιληπτές χρωματίζοντας σε κάθε μεταβλητή όπου υπάρχει διαφορά από συστάδα σε συστάδα, κάτι που είναι πολύ χρήσιμο και διευκολύνει τον χρήστη να αντιληφθεί οπτικά πιο εύκολα τις διαφορές όπως φαίνεται και στην εικόνα 5.9 (πράσινο χρώμα). Το πόσες εγγραφές ανήκουν σε κάθε συστάδα θα το δούμε ανοίγοντας το "*Attribute Statistics*" (διπλό κλικ) και επιλέγοντας την μεταβλητή Cluster. Μετά πατάμε Save Graph και αποθηκεύεται το γράφημα. Κάθε γράφημα μπορεί να αποθηκευτεί με τον ίδιο τρόπο. Στην εικόνα 5.11 φαίνεται η κατανομή των εγγραφών.



Εικόνα 5.11: Η κατανομή των 3 συστάδων

Στην πρώτη συστάδα όπως βλέπουμε από το "*Data Table*" το προφίλ του πελάτη είναι το εξής: Άντρας, παντρεμένος με τουλάχιστον 1 παιδί, 34 χρονών, με μισθό 1653 δολάρια, αποταμιεύσεις 9498,1 δολάρια, έχει μηνιαία αξία τραπεζικών κονδυλίων 1559 δολάρια, κατοικεί στην πολιτεία της Νέας Υόρκης, είναι ιδιοκτήτης 1 κατοικίας ,έχει αυτοκίνητο , 1 υποθήκη και είναι 3 χρόνια πελάτης της τράπεζας.

Στην δεύτερη συστάδα το προφίλ του πελάτη είναι ως εξής: Άντρας, ανύπαντρος χωρίς παιδιά, 34 χρονών με μισθό 250 δολάρια, αποταμιεύσεις 1841,9 δολάρια, έχει μηνιαία αξία τραπεζικών κονδυλίων 701 δολάρια, κατοικεί στην πολιτεία της Νέας Υόρκης, δεν είναι ιδιοκτήτης κατοικίας και ως εκ τούτου δεν έχει υποθήκη, έχει αυτοκίνητο και είναι 3 χρόνια πελάτης της τράπεζας.

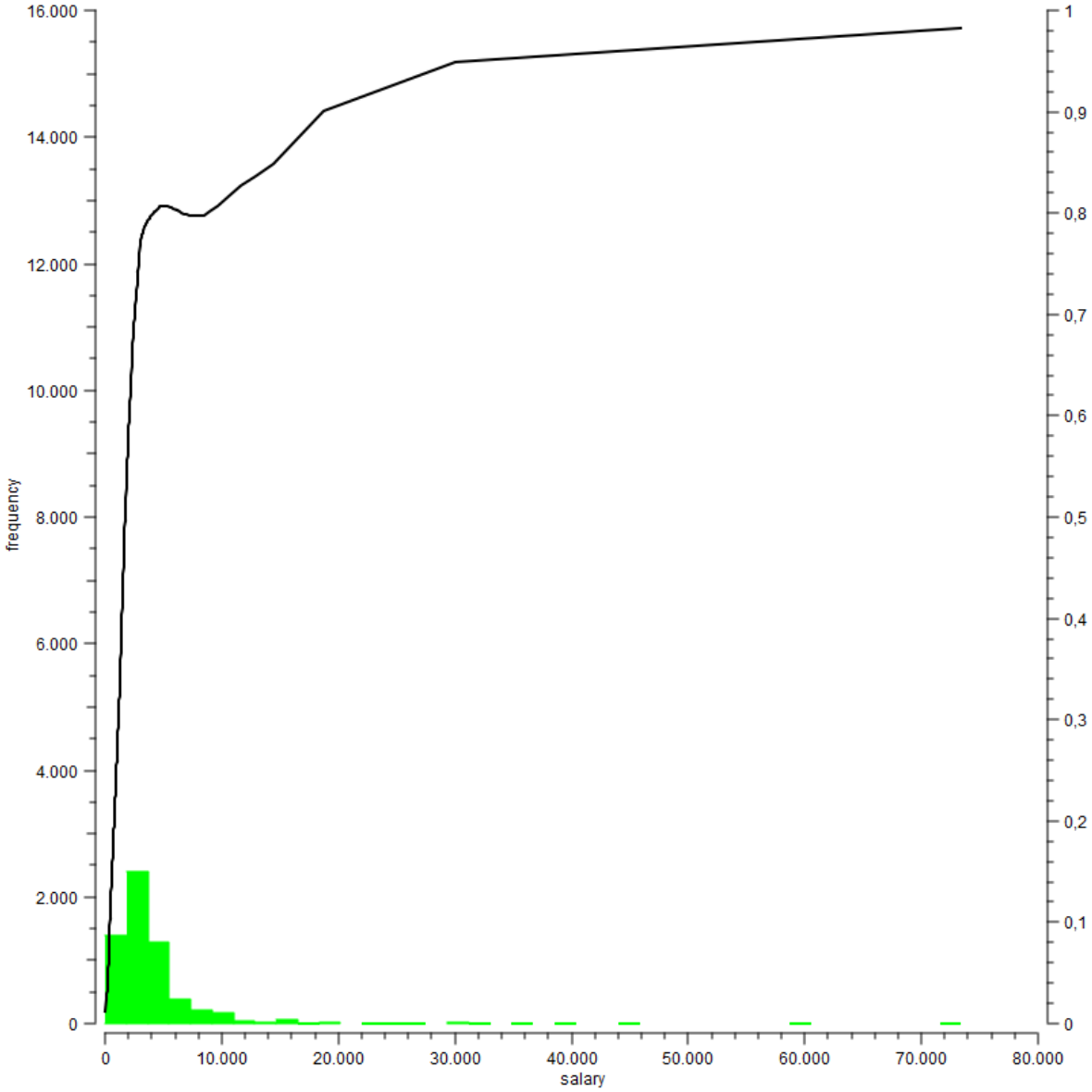
Στην τρίτη συστάδα το προφίλ του πελάτη είναι: Γυναίκα με διαζύγιο, με τουλάχιστον 1 παιδί, 45 ετών, με μισθό 3813 δολάρια, αποταμιεύσεις 70610,1 δολάρια, έχει μηνιαία αξία τραπεζικών κονδυλίων 5061 δολάρια, κατοικεί στην πολιτεία της Νέας Υόρκης, είναι ιδιοκτήτρια 1 κατοικίας, έχει αυτοκίνητο, 1 υποθήκη, και είναι ένα χρόνο πελάτης της τράπεζας.

Όπως είναι πολύ προφανές αυτό που χαρακτηρίζει τις 3 συστάδες είναι η οικονομική επιφάνεια των πελατών που ανήκουν σε αυτές. Η **συστάδα 2** με μέσο μισθό 250 περιέχει τους ανέργους και τους χαμηλόμισθους που έχουν μικρές αποταμιεύσεις και δεν έχουν κάνει ακόμα οικογένεια. Οι συναλλαγές που έχουν με την τράπεζα είναι πολύ μικρές καθώς και τα χρήματα που τους ανήκουν είναι πολύ λίγα. Στην **συστάδα 1** ανήκουν οι πελάτες με τα μεσαία εισοδήματα που έχουν κάποιες αποταμιεύσεις, έχουν κάνει οικογένεια και έχουν ικανοποιητικές συναλλαγές με την τράπεζα καθώς έχουν ακίνητο και πληρώνουν υποθήκη ενώ και ο μισθός τους τους επιτρέπει να κάνουν κάποιες μικρές επενδύσεις σε ομόλογα και σε μετοχές. Στην **συστάδα 3** που είναι και οι περισσότεροι πελάτες ανήκει η «ελίτ» των πελατών της τράπεζας με τον μεγαλύτερο μισθό τις πολύ υψηλές αποταμιεύσεις και τις πολλές επενδύσεις σε μετοχές και ομόλογα. Οι συναλλαγές που κάνουν με την τράπεζα είναι κατά πολύ υψηλότερη από τους πελάτες των άλλων συστάδων.

Το πώς μπορούμε να δούμε σχηματικά τις κατανομές είναι πολύ εύκολο να το κάνουμε μέσα από το widget "*Distributions*". Αριστερά είναι οι επιλογές και οι ρυθμίσεις που μπορούμε να κάνουμε και δεξιά είναι το γράφημα που προκύπτει από κάθε επιλογή που κάνουμε. Στην επιλογή Variable επιλέγουμε την μεταβλητή που θέλουμε να εμφανιστεί και στο Displayed outcomes επιλέγουμε από ποια απ τις συστάδες θέλουμε να δούμε την κατανομή. Μπορούμε να συνδυάσουμε τις συστάδες και κάθε μια συστάδα έχει διαφορετικό χρώμα στο γράφημα έτσι ώστε να φαίνεται ποια κατανομή ανήκει σε κάθε μια απ αυτές. Μπορείς να προσαρμόσεις τις ονομασίες το x και του y άξονα αν και ο y είναι μόνιμα η συχνότητα κι αυτό δεν αλλάζει. Τέλος έχει και την γραμμή Probability που δείχνει πόσο πιθανό είναι σε κάθε τιμή να ανήκει κάποιος πελάτης σε συγκεκριμένη συστάδα. Για παράδειγμα, στην εικόνα 5.12 βλέπουμε την κατανομή των πελατών που ανήκουν στην συστάδα 3 και η γραμμή πιθανοτήτων μας δείχνει πως όσο υψηλότερος είναι ο μισθός μιας εγγραφής τόσο πιθανότερο είναι να ανήκει στην τρίτη συστάδα.

Πολύ χρήσιμα συμπεράσματα μπορούν να βγούνε και από τα γραφήματα με μορφή διασποράς και το Orange έχει ένα widget ονόματι "*Scatterplo*t" στο μενού Visualize που είναι πολύ χρήσιμο. Βάζοντας ως είσοδο το αποτέλεσμα του "*k-means clustering*" γίνεται δυνατός ο συνδυασμός δυο μεταβλητών και με χρώμα να βλέπεις κάποιες κατανομές. Όπως και στο "*Distributions*" έτσι και στο

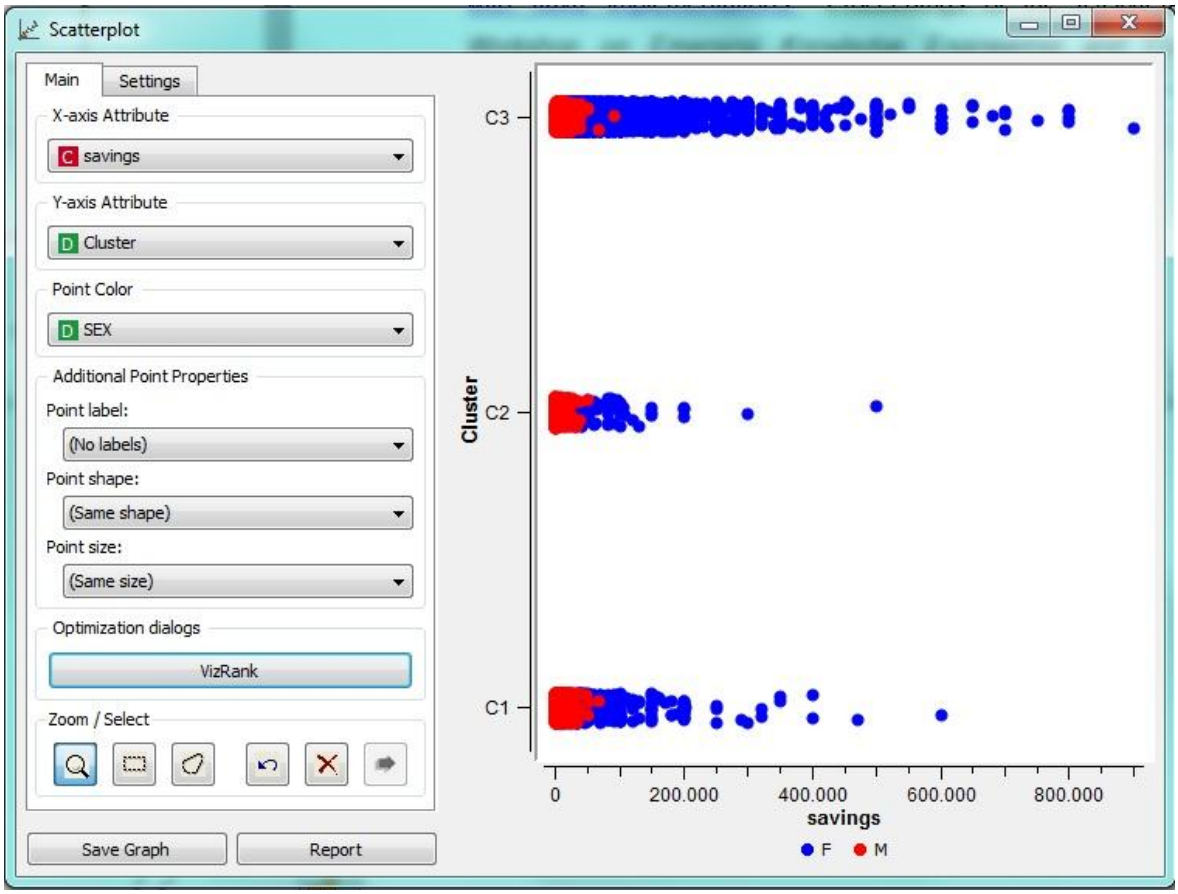
“*Scatterplot*” αριστερά είναι οι επιλογές και δεξιά είναι στο γράφημα. Στις επιλογές επιλέγονται οι μεταβλητές για τους άξονες x και y και ο χρωματισμός για μια τρίτη μεταβλητή. Για παράδειγμα, στην εικόνα 5.13 έχει επιλεγεί ως άξονας x οι αποταμιεύσεις, ως y οι 3 συστάδες και το φύλλο να είναι χρωματισμένο, δηλαδή οι άντρες και οι γυναίκες να έχουν διαφορετικό χρώμα. Κατά την αποθήκευση του διαγράμματος μέσω του Save Graph διαπιστώθηκε ότι το τα υπομνήματα αποθηκεύτηκαν λανθασμένα, οπότε προτιμήθηκε η αποθήκευση με Print Screen.



Εικόνα 5.12: Η κατανομή μισθών της συστάδας 3 με την γραμμή πιθανοτήτων

Όπως παρατηρούμε στην εικόνα 5.13, οι άντρες έχουν τις χαμηλότερες αποταμιεύσεις ενώ οι γυναίκες έχουν τις υψηλότερες και μάλιστα αυτό ισχύει και στις 3 συστάδες. Μπορεί να γίνει οποιοσδήποτε συνδυασμός μεταβλητών και

χρωματισμών κάνοντας έτσι το “Scatterplot” ένα χρησιμότατο widget που παρέχει το Orange.



Εικόνα 5.13: Το παράθυρο Scatterplot στο Orange

5.3 Rapidminer

5.3.1 Επιλογή μεταβλητών

Ανοίγοντας το Rapidminer επιλέγουμε New και στην ερώτηση αν θέλουμε να ορίσουμε κάποιο φάκελο ως Repository επιλέγουμε όχι γιατί δεν χρειάζεται στο συγκεκριμένο παράδειγμα. Βλέποντας το περιβάλλον του προγράμματος, αριστερά βρίσκονται ομαδοποιημένοι όλοι οι τελεστές, στο κέντρο βρίσκεται η διεργασία, δεξιά η παραμετροποίηση του εκάστοτε επιλεγμένου τελεστή και στο κάτω μέρος το Log και τα errors που μπορεί να εμφανιστούνε.

Επόμενο βήμα είναι η εισαγωγή του αρχείου στην διεργασία. Από τους τελεστές ανοίγουμε τον φάκελο Import μετά Data και επιλέγουμε το “*Read CSV*”

και το σύρουμε στο παράθυρο της διεργασίας. Πατώντας πάνω στον τελεστή αυτό δεξιά βλέπουμε τις παραμέτρους του. Πατάμε στο κουμπί "*Import Configuration Wizard"* και ανοίγει ένα νέο παράθυρο που μας λέει να του ορίσουμε ποιο είναι το csv αρχείο που θέλουμε να εισάγουμε. Το επιλέγουμε και πατάμε Next. Στην συνέχεια στο 2^{ο} βήμα του ορίζουμε το κόμμα ως διαχωριστικό των στηλών και κάτω μας έχει ένα Preview του πως βλέπει το αρχείο. Συνεχίζουμε στο 3^{ο} βήμα όπου και ορίζουμε την πρώτη σειρά ως το όνομα της κάθε μεταβλητής και πατάμε Next. Στο 4^{ο} βήμα επιλέγουμε τον τύπο της κάθε μεταβλητής καθώς και αποεπιλέγουμε όποιες μεταβλητές δεν θέλουμε να πάρουν μέρος στην διεργασία μας. Όπως έχουμε προαποφασίσει το Client_id και το profession θα βγούνε. Έπειτα πατάμε το κουμπί "Guess value types" που ορίζει αυτόματα τον τύπο της κάθε μεταβλητής. Αφού κάνουμε έναν απαραίτητο έλεγχο αν επέλεξε σωστούς τύπους για τις μεταβλητές πατάμε Finish.

5.3.2 Μετασχηματισμός μεταβλητών

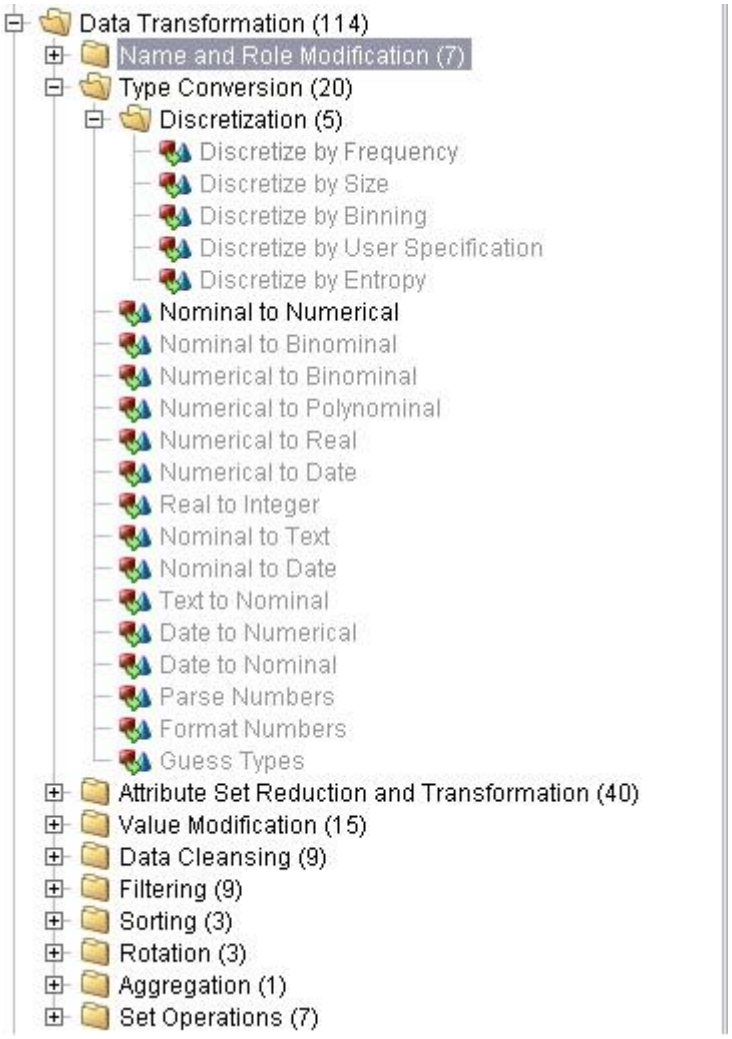
Εφόσον ορίσαμε μέσα από τον οδηγό που παρέχει το πρόγραμμα με μεγάλη ευκολία και ταχύτητα τους τύπους των μεταβλητών ας δούμε τι άλλες δυνατότητες έχει το πρόγραμμα ως προς το μετασχηματισμό των μεταβλητών.

Στους τελεστές στο αριστερό κομμάτι του παραθύρου, στον φάκελο Data Transformation υπάρχουν συνολικά 114 διαφορετικοί τελεστές για τον μετασχηματισμό μεταβλητών και μάλιστα ομαδοποιημένους. Χωρίζονται σε 10 κατηγορίες που είναι οι εξής:

- Name and Role Modification: Περιέχει 7 τελεστές και οι βασικότεροι είναι η μετονομασία, ανταλλαγή ρόλων μεταξύ δυο μεταβλητών ή το να τεθεί νέος ρόλος σε κάποια μεταβλητή
- Type Conversion: Περιέχει 20 τελεστές εκ των οποίων οι 5 είναι για το Discretization είτε by Entropy είτε by Size είτε by Frequency κτλ. Οι υπόλοιποι 15 αφορούν μετασχηματισμούς τύπων δηλαδή Nominal to Numerical, Nominal to Binominal κτλ. (Εικόνα 5.14)
- Attribute Set Reduction and Transformation: Περιέχει 40 τελεστές, 20 εκ των οποίων για δημιουργία νέων μεταβλητών, 7 για μαθηματικό

μετασχηματισμό και 13 για επιλογή (τυχαία, by weight, remove useless attributes κτλ)

- Value Modification: Περιέχει 15 τελεστές για τροποποίηση αξιών στις μεταβλητές μεταξύ των οποίων και το Normalize (κανονικοποίηση).
- Data Cleansing: Περιέχει 9 τελεστές, 4 εκ των οποίων για Outlier Detection και 5 για Replace Missing Values.
- Filtering: Περιέχει 9 τελεστές για φιλτράρισμα ή δειγματισμό δεδομένων
- Sorting: Περιέχει 3 τελεστές για ταξινόμηση δεδομένων
- Rotation: Περιέχει 3 τελεστές για Pivot
- Aggregation: Περιέχει τον τελεστή Aggregate
- Set Operations: Περιέχει 7 τελεστές για συνδυασμό από example set
-



Εικόνα 5.14: Μια άποψη των τελεστών του Type Conversion για τον μετασχηματισμό μεταβλητών

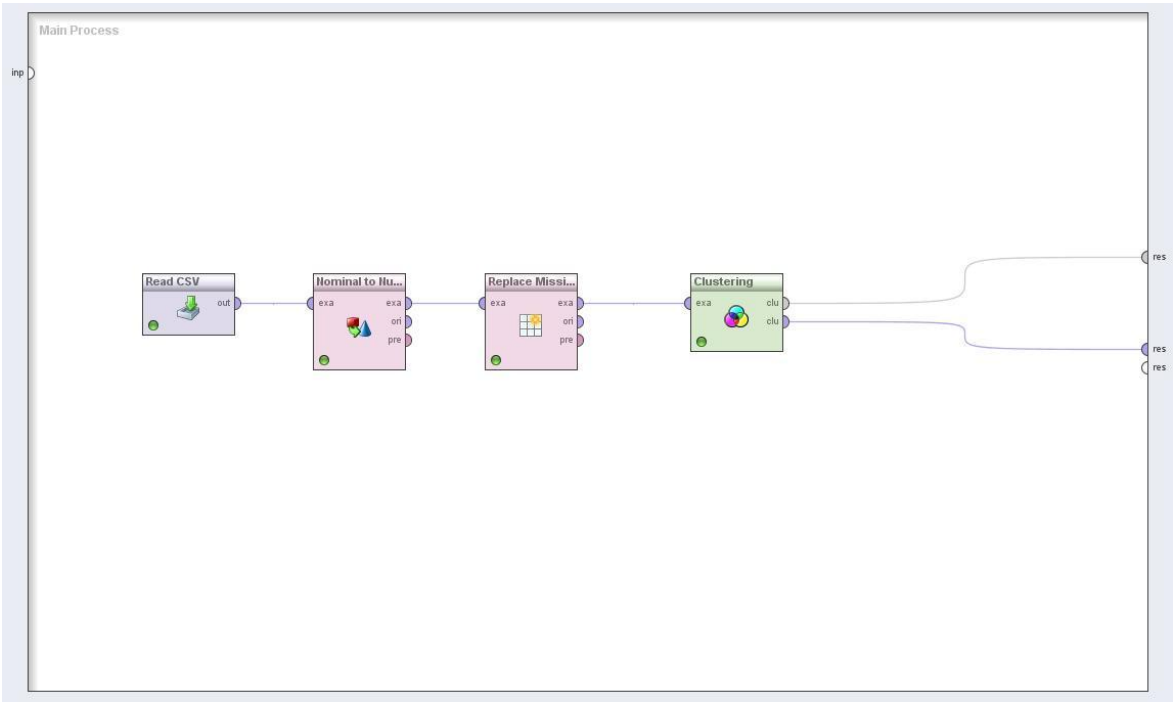
Όταν πατάμε σε κάθε ένα από τους τελεστές, τότε κάτω δεξιά στο παράθυρο του προγράμματος ανοίγει η περιγραφή για το τι ακριβώς κάνει ο κάθε τελεστής κάνοντας πολύ πιο εύκολη την επιλογή του κατάλληλου τελεστή για την εκάστοτε διεργασία μας.

5.3.3 Προσδιορισμός παραμέτρων

Ήρθε η ώρα λοιπόν να επιλέξουμε τον kmeans και να ορίσουμε τις παραμέτρους του αλγορίθμου. Το Rapidminer έχει μια ιδιαιτερότητα εδώ. Έχει τον δικό της αλγόριθμο kmeans ο οποίος όμως δεν μπορεί να διαχειριστεί μη αριθμητικές μεταβλητές και τον kmeans του Weka που είναι πανομοιότυπος με τον αλγόριθμο στο Weka. Τα αποτελέσματα όμως σε αυτόν είναι «φτωχά». Παρακάτω θα δείξουμε σε κάθε μια από τις 2 περιπτώσεις τι αποτελέσματα και τι προβλήματα συναντήσαμε αναλυτικά.

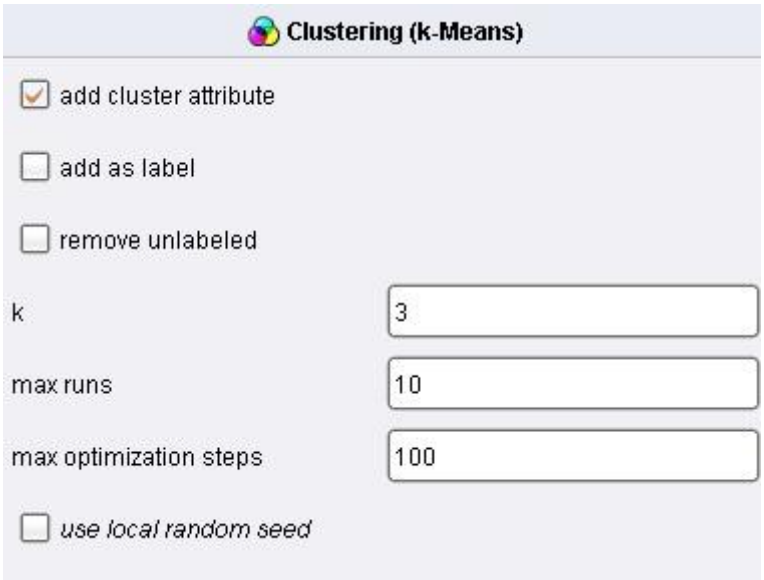
Ξεκινάμε με τον αλγόριθμο του Rapidminer. Κάνουμε δεξί κλικ στην διεργασία μετά πάμε στον φάκελο Modeling και μετά Clustering and Segmentation και επιλέγουμε τον αλγόριθμο kmeans. Ενώνουμε το output του "*Read CSV*" με το input του "*kmeans*". Στο πεδίο που εμφανίζει τα προβλήματα λαμβάνουμε το εξής μήνυμα "k-Means cannot handle binominal attributes" και προτείνει ως λύση να μετατρέψουμε τις μεταβλητές σε numerical. Διπλοπατώντας πάνω στην γραμμή που εμφανιζόταν το μήνυμα, αυτόματα εισάγει έναν νέο τελεστή ανάμεσα στον "*kmeans*" και στον "*Read CSV*" που λέγεται "*Nominal to Numerical*". Ενώνουμε λοιπόν τις εξόδους του "*kmeans*" (data και metadata) με τα results για να πάρουμε αποτέλεσμα. Όταν πατάμε να τρέξει βγαίνει ένα παράθυρο λάθους κατά την εκτέλεση που μας ενημερώνει ότι υπάρχουν στα δεδομένα μας missing values και δεν μπορεί να τα διαχειριστεί ο kmeans. Προτείνει να εισάγουμε έναν τελεστή για να αντικαταστήσει τα κενά. Κάνουμε δεξί κλικ στην διεργασία για να εισάγουμε τον νέο τελεστή και επιλέγουμε από τον φάκελο Data Transformation και μετά Data Cleansing τον "*Replace Missing Values*". Στην συνέχεια το βάζουμε ανάμεσα στο "*Read CSV*" και τον "*kmeans*". Επιλέγοντας αυτόν τον τελεστή, στις παραμέτρους βλέπουμε ότι αντικαθιστά τα κενά με τον μέσο όρο (default) και έχει επιλογές min,max,zero κτλ. Έτσι η τελική μορφή της διεργασίας με τον αλγόριθμο του

Rapidminer είναι όπως φαίνεται στην εικόνα 5.15. Υπάρχει η επιλογή export process που επιτρέπει την εξαγωγή της διεργασίας ως εικόνα,pdf κα.



Εικόνα 5.15: Η διεργασία στο Rapidminer με τον αλγόριθμο kmeans του προγράμματος.

Κάνοντας κλικ στον "*kmeans*" δεξιά βλέπουμε την παραμετροποίησή του. Στην εικόνα 5.16 είναι οι επιλογές που υπάρχουν.

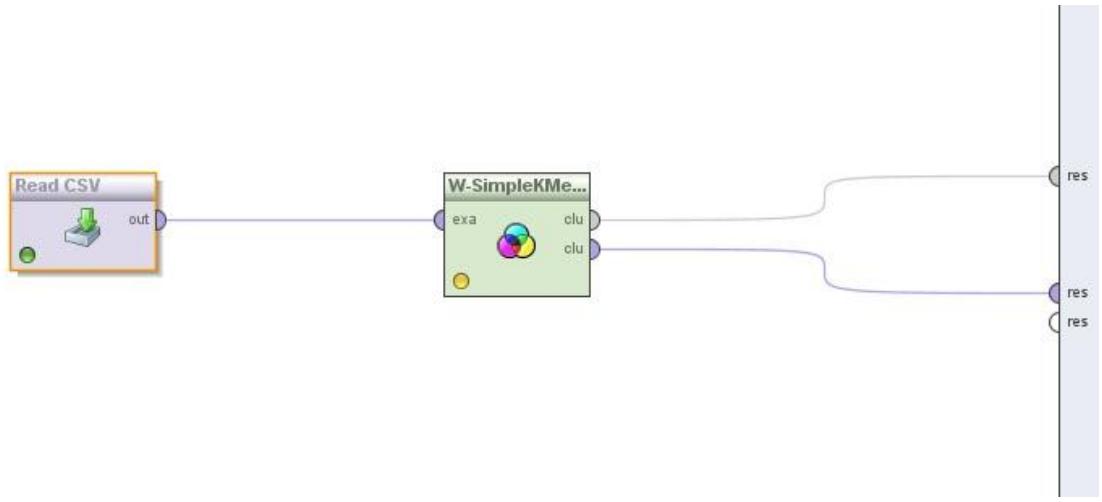


Εικόνα 5.16: Οι παράμετροι στον αλγόριθμο kmeans του προγράμματος Rapidminer

Θέλουμε να εισάγει μια νέα μεταβλητή που να δηλώνει την συστάδα της κάθε εγγραφής γι αυτό και είναι επιλεγμένο το "*add cluster attribute*". Το "*add as label*" το επιλέγουμε όταν θέλουμε το cluster attribute ναι είναι σε μορφή ετικέτας, με το "*remove unlabeled*" αφαιρεί τα παραδείγματα που δεν έχουν ετικέτα. Το k είναι ο

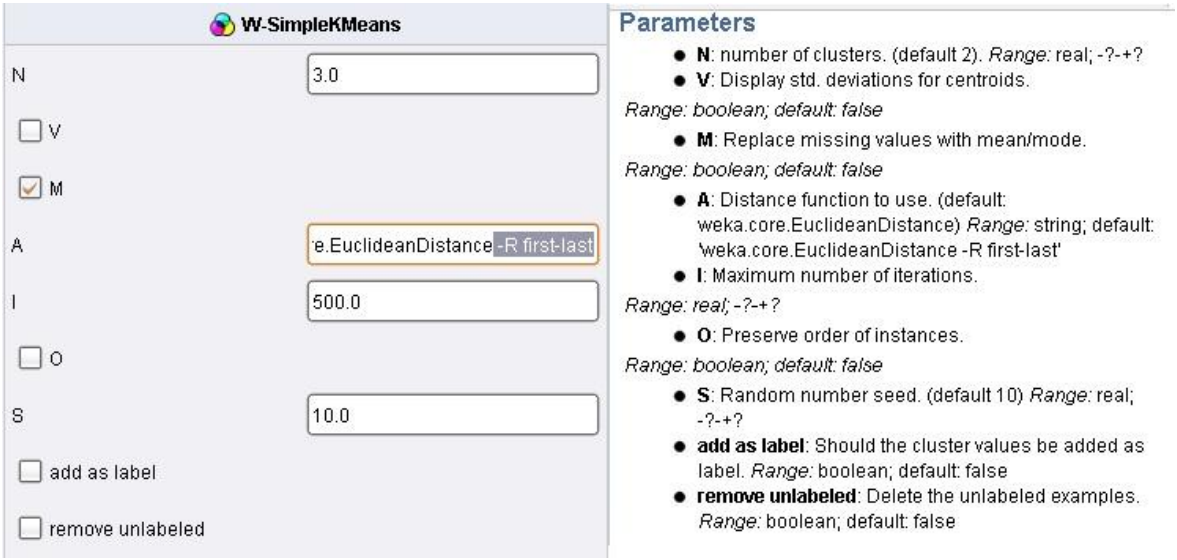
αριθμός των συστάδων που προκαθορίζουμε, το "*max runs*" είναι ο μέγιστος αριθμός εκτελέσεων του αλγορίθμου με τυχαία εκκίνηση. Το "*max optimization steps*" είναι ο μέγιστος αριθμός επαναλήψεων που επιτρέπεται σε μια εκτέλεση του αλγορίθμου. Με το "*use local random seed*" επιλέγουμε αν θέλουμε να του ορίσουμε εμείς τον αριθμό του seed.

Τα αποτελέσματα της εκτέλεσης του αλγορίθμου θα τα δούμε στην επόμενη ενότητα. Τώρα θα περιγραφεί η παραμετροποίηση καθώς και η διαδικασία για τον αλγόριθμο kmeans του Weka που υπάρχει στον Rapidminer. Έχοντας ήδη έτοιμο το "*Read CSV*" κάνουμε δεξί κλικ στην διεργασία new operator->Modeling->Clustering and Segmentation->Weka και επιλέγουμε τον "*W-Simplekmeans*". Αντίθετα με τον αλγόριθμο του Rapidminer στον συγκεκριμένο δεν αντιμετωπίζουμε κάποιο πρόβλημα σε σχέση με τα nominal όπως είχε συμβεί και με τα άλλα 2 προγράμματα. Στην εικόνα 5.17 βλέπουμε την απλή διεργασία με μόνη είσοδο το αρχείο χωρίς παρεμβολές.



Εικόνα 5.17: Η διεργασία με τον W-Simplekmeans στο Rapidminer

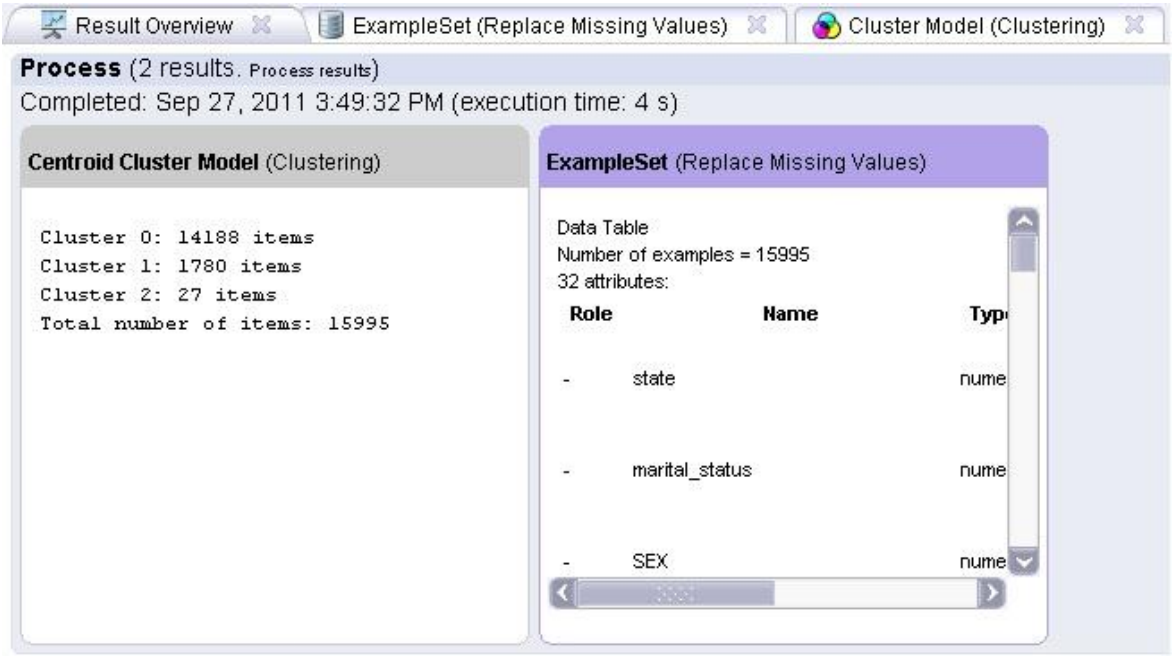
Ως προς την παραμετροποίηση είναι σχεδόν ίδια με την παραμετροποίηση του Weka και είναι λογικό καθώς είναι ο ίδιος αλγόριθμος. Στην εικόνα 5.18 βλέπουμε τις παραμέτρους που μπορούμε να ρυθμίσουμε και δεξιά τις περιγραφές των ρυθμίσεων που υπάρχουν κάτω δεξιά στο κεντρικό παράθυρο του Rapidminer. Το N είναι ο αριθμός των συστάδων, το V το επιλέγουμε όταν θέλουμε να δούμε τις τυπικές αποκλίσεις, το M είναι για να αντικαταστήσει τις κενές αξίες με τον μέσο όρο, το A για την μονάδα μέτρησης αποστάσεων, το I για τον μέγιστο αριθμό επαναλήψεων, το O αν θέλουμε να διατηρηθεί η σειρά των εγγραφών και το S είναι ο τυχαίος αριθμός seed.



Εικόνα 5.18: Οι παράμετροι του W-Simplekmeans στο Rapidminer και οι περιγραφές που παρέχει

5.3.4 Εφαρμογή και αποτελέσματα

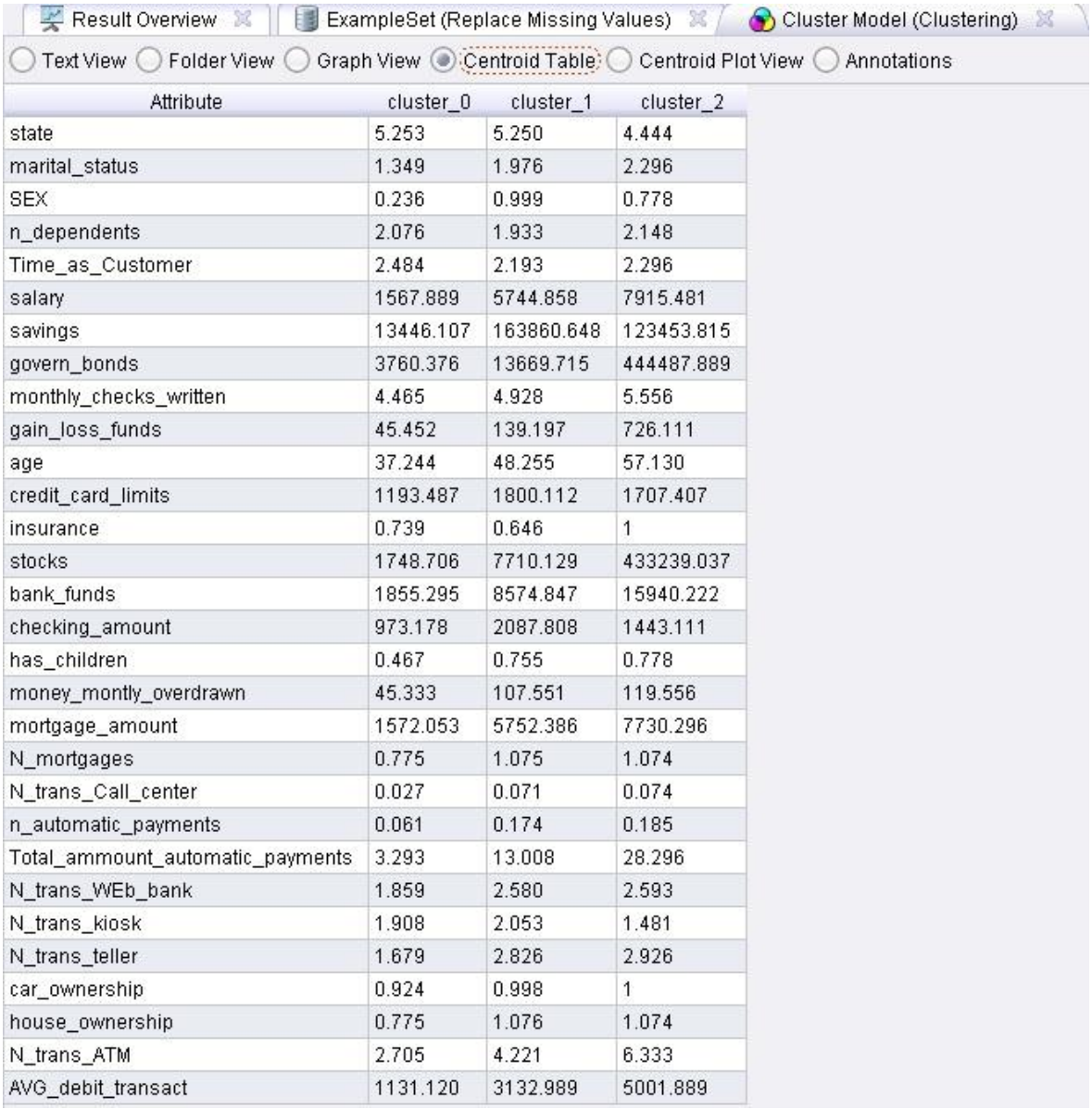
Τώρα θα εκτελέσουμε τον αλγόριθμο “*kmeans*” του Rapidminer με τις παραμέτρους που αναλύθηκαν στην προηγούμενη ενότητα. Πατάμε το κουμπί της εκτέλεσης (μοιάζει με το play ενός cd) ,μας ρωτάει αν θέλουμε να αποθηκευτεί η διεργασία , μετά μας ενημερώνει ότι έχουν βγει νέα αποτελέσματα και μας ρωτάει αν θέλουμε να περάσουμε στο result perspective όπου βλέπουμε αναλυτικά τα αποτελέσματα για οποιονδήποτε αλγόριθμο έχουμε. Επιλέγουμε λοιπόν ναι. Στην καρτέλα “Result Overview” βλέπουμε πότε βγήκαν τα αποτελέσματα (σε περίπτωση που έχουμε πολλά), πόσος ήταν ο χρόνος εκτέλεσης (4 δευτερόλεπτα στην περίπτωση αυτή) και κάνοντας κλικ στο αποτέλεσμα βλέπουμε τα αποτελέσματα όπως είναι στην εικόνα 5.19. Βλέπουμε ότι η πρώτη συστάδα έχει την μερίδα του λέοντος των στοιχείων (14188), η δεύτερη έχει 1780 και η τρίτη μόνο 27! Στην καρτέλα “ExampleSet” βλέπουμε η κάθε εγγραφή σε ποια συστάδα πήγε και αναλυτικά τις μεταβλητές και τις αξίες τους. Υπάρχει η δυνατότητα να δει κάποιος της κατανομές των μεταβλητών με έναν Plotter διαλέγοντας μέσα από διαφόρων ειδών γραφήματα. Η πιο ενδιαφέρουσα καρτέλα είναι η “Cluster Model” όπου υπάρχει και το centroid table με τους μέσους όρους της κάθε συστάδας. Στην εικόνα 5.20 παρατίθεται ο συγκεκριμένος πίνακας όπως παρέχεται από το πρόγραμμα.



Εικόνα 5.19: Η καρτέλα "Result Overview" των αποτελεσμάτων στο Rapidminer

Όπως βλέπουμε πλέον η μεταβλητή state δεν έχει κανένα νόημα ως αριθμητική, όπως και το marital _status. Στην μεταβλητή SEX επειδή μεταφράστηκε το αρσενικό ως 0 και το θηλυκό ως 1 μπορεί να χρησιμοποιηθεί και να βγει κάποιο συμπέρασμα.

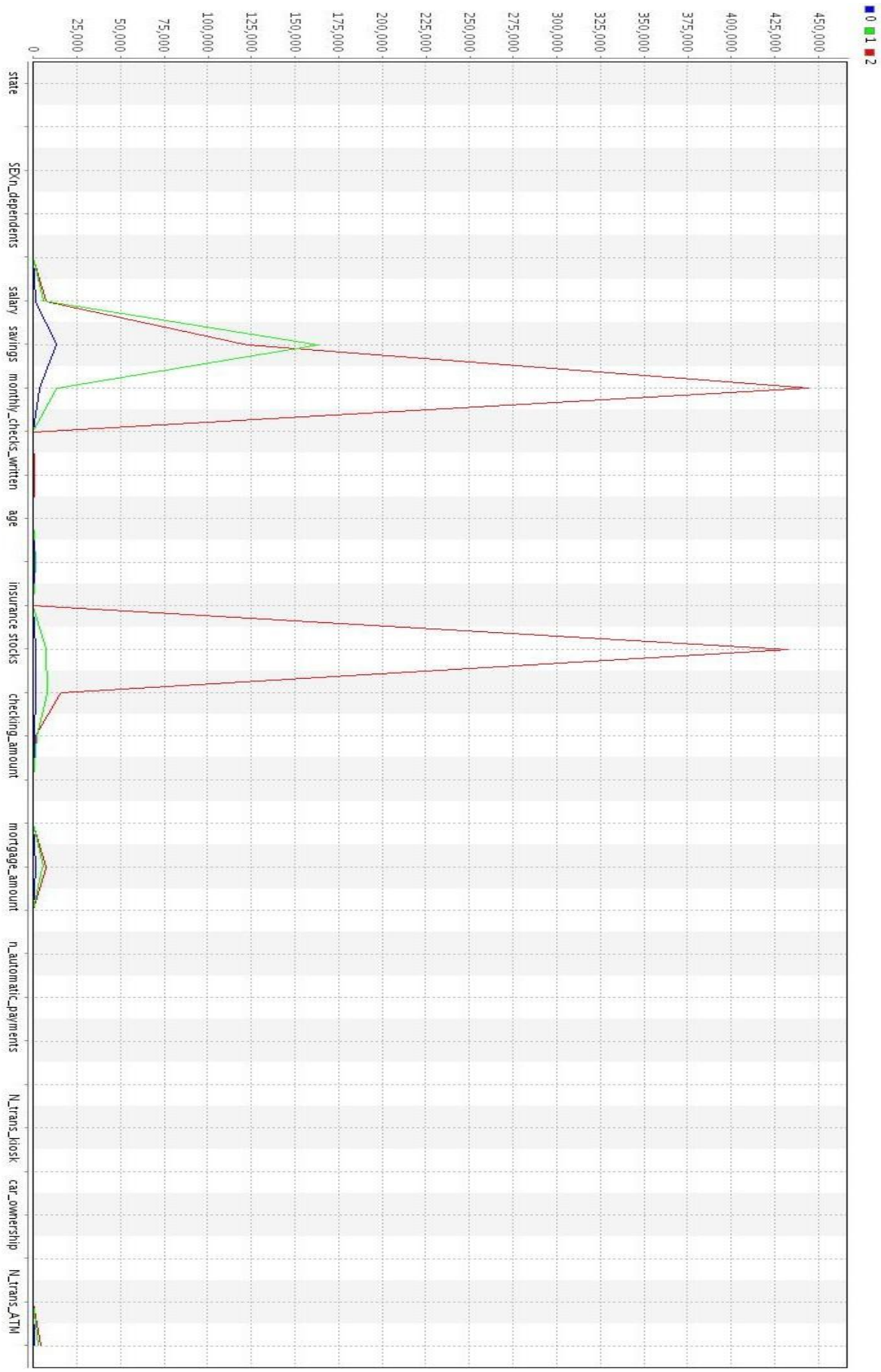
Ας δούμε τώρα τα προφίλ των τριών συστάδων. Η **συστάδα 0** ,που έχει και τα περισσότερα στοιχεία, βλέπουμε ότι αποτελείται κατά 76% από άντρες που έχουν 2 περίπου εξαρτώμενους, είναι 2,5 χρόνια πελάτες της τράπεζας, έχει μισθό 1567 δολάρια και καταθέσεις 13446 δολάρια. Είναι 37 χρονών, έχει ασφάλεια με 73% πιθανότητα, 46% πιθανότητα να έχει ένα τουλάχιστον παιδί και έχει αυτοκίνητο με 91% πιθανότητα. Η **συστάδα 1** αποτελείται εξ ολοκλήρου από γυναίκες (0,999 στην μεταβλητή SEX), έχουν 2 περίπου εξαρτώμενους είναι 2 χρόνια πελάτες της τράπεζας έχουν μισθό 5744 δολάρια και καταθέσεις 163860 δολάρια. Είναι 48 χρονών, έχει ασφάλεια με πιθανότητα 65%, 75% πιθανότητα να έχει ένα τουλάχιστον παιδί και έχει σίγουρα αυτοκίνητο (0,998 car ownership). Η **συστάδα 2** αποτελείται κατά 78% από γυναίκες που έχουν 2 εξαρτώμενους ,είναι 2 χρόνια πελάτες της τράπεζας, έχει μισθό 7915 δολάρια και καταθέσεις 123453 δολάρια. Είναι 57 χρονών έχει σίγουρα ασφάλεια, έχει 78% πιθανότητα να έχει ένα τουλάχιστον παιδί και έχει σίγουρα αυτοκίνητο.



Attribute	cluster_0	cluster_1	cluster_2
state	5.253	5.250	4.444
marital_status	1.349	1.976	2.296
SEX	0.236	0.999	0.778
n_dependents	2.076	1.933	2.148
Time_as_Customer	2.484	2.193	2.296
salary	1567.889	5744.858	7915.481
savings	13446.107	163860.648	123453.815
govern_bonds	3760.376	13669.715	444487.889
monthly_checks_written	4.465	4.928	5.556
gain_loss_funds	45.452	139.197	726.111
age	37.244	48.255	57.130
credit_card_limits	1193.487	1800.112	1707.407
insurance	0.739	0.646	1
stocks	1748.706	7710.129	433239.037
bank_funds	1855.295	8574.847	15940.222
checking_amount	973.178	2087.808	1443.111
has_children	0.467	0.755	0.778
money_montly_overdrawn	45.333	107.551	119.556
mortgage_amount	1572.053	5752.386	7730.296
N_mortgages	0.775	1.075	1.074
N_trans_Call_center	0.027	0.071	0.074
n_automatic_payments	0.061	0.174	0.185
Total_ammount_automatic_payments	3.293	13.008	28.296
N_trans_WEb_bank	1.859	2.580	2.593
N_trans_kiosk	1.908	2.053	1.481
N_trans_teller	1.679	2.826	2.926
car_ownership	0.924	0.998	1
house_ownership	0.775	1.076	1.074
N_trans_ATM	2.705	4.221	6.333
AVG_debit_transact	1131.120	3132.989	5001.889

Εικόνα 5.20: Ο πίνακας με τα centroids των συστάδων στα αποτελέσματα του “kmeans” στο Rapidminer

Παρατηρούμε ότι υπάρχει διαφορά στους μισθούς κλιμακωτά, όπως επίσης στις ηλικίες και στα κρατικά ομόλογα όπως και στην μηνιαία αξία τραπεζικών κονδυλίων(bank funds) που είναι και οι μεγαλύτερες διαφορές από συστάδα σε συστάδα. Για την ακόμα πιο εύκολη παρατήρηση των διαφορών στις συστάδες ,έτσι όπως διαχωρίστηκαν με τον αλγόριθμο, υπάρχει ένα γράφημα στο Centroid Plot View. Σ αυτό το γράφημα εμφανίζονται τα δεδομένα του πίνακα στην εικόνα 5.20 σχηματικά. Το σχήμα αυτό φαίνεται στην εικόνα 5.21.



Εικόνα 5.21: Το centroid plot view στα αποτελέσματα του Rapidminer.

Αφού είδαμε την εκτέλεση και τα αποτελέσματα του αλγορίθμου “*kmeans*” ας δούμε τι αποτελέσματα παίρνουμε κατά την εκτέλεση του “*W-Simplekmeans*” έτσι όπως ορίσαμε τις παραμέτρους στην προηγούμενη ενότητα. Υπάρχουν οι ίδιες καρτέλες όπως και στον “*kmeans*”. Στην συστάδα 0 υπάρχουν 5927 στοιχεία , στην συστάδα 1 4724 στοιχεία και στην συστάδα 2 5344 στοιχεία. Μια πρώτη παρατήρηση είναι ότι ενώ είναι ακριβώς ο ίδιος αλγόριθμος με του Weka, οι συστάδες στον Rapidminer δεν έχουν τα ίδια στοιχεία με τα αποτελέσματα των συστάδων του Weka όπως είδαμε και στην ενότητα της εκτέλεσης του αλγορίθμου kmeans στο Weka. Στην καρτέλα Cluster Model των αποτελεσμάτων απουσιάζουν και το Centroids Table και το Centroid Plot View. Αυτό σημαίνει ότι δεν μπορούμε να έχουμε τους μέσους όρους των μεταβλητών ανά συστάδα που είναι ο πιο σημαντικός πίνακας αποτελεσμάτων για την εξαγωγή συμπερασμάτων. Έχουμε δηλαδή μόνο τον πίνακα με όλες τις εγγραφές και σε ποια συστάδα ανήκουν αλλά όχι συγκεντρωτικά αποτελέσματα. Ως εκ τούτου είναι αδύνατον να σκιαγραφήσουμε το προφίλ της κάθε συστάδας και να αξιολογήσουμε περαιτέρω τα αποτελέσματα.

ΚΕΦΑΛΑΙΟ 6
Σύγκριση των Λογισμικών Πακέτων

6.1 Γενικά στοιχεία

Η πρώτη εντύπωση που αποκομίζεται χρησιμοποιώντας για πρώτη φορά ένα εργαλείο, πολλές φορές, είναι αποθαρρυντικό για την περαιτέρω ενασχόλησή με αυτό. Η βασική αιτιολογία έχει να κάνει με το κατά πόσο είναι φιλικό στον χρήστη, ειδικά εάν ο χρήστης είναι νέος στην εξόρυξη πληροφορίας και επιθυμεί να πειραματιστεί με αυτό. Τα λογισμικά πακέτα που πραγματεύεται η παρούσα εργασία έχουν διαφορετικές επιδόσεις σε αυτόν τον τομέα.

Σε ότι αφορά στο **Weka**, αυτό προσφέρει κάποιες powerpoint παρουσιάσεις και manual αλλά δεν παρέχει κάποιο tutorial για τον νέο χρήστη. Ανοίγοντας την εφαρμογή και δοκιμάζοντας από περιβάλλον σε περιβάλλον (explorer, knowledge flow κτλ) να περιηγηθεί ο νέος χρήστης, του δημιουργεί την εντύπωση ενός εργαλείου που σε παραπέμπει στις αρχές της δεκαετίας του 2000. Η πρώτη εικόνα λοιπόν, έχει ανασταλτικό χαρακτήρα στην επιλογή του Weka ως εργαλείου για το data mining. Η εικόνα του σε σχέση με τα άλλα δυο που θα συζητήσουμε παρακάτω δείχνει ένα εργαλείο μη σύγχρονο.

Σε ότι αφορά στο **Orange,** η ιστοσελίδα του προσφέρει σχεδόν καθημερινή νέα «έκδοση» του προγράμματος που δείχνει ότι είναι ένα συνεχώς εξελισσόμενο εργαλείο σε αντίθεση με το Weka. Υπάρχουν manuals και για αρχαρίους στο http://orange.biolab.si/doc/ofb/ αλλά και forum που μπορεί κάποιος να θέσει κάποιες ερωτήσεις για το εργαλείο και να λάβει σχετικά σύντομα απάντηση από άλλους χρήστες ή ακόμα και από τους προγραμματιστές που το αναπτύσσουν. Το γεγονός ότι το περιβάλλον του Orange, σε αντίθεση με το πιο διαδεδομένο περιβάλλον του Weka το Explorer, παρέχει την δυνατότητα εισαγωγής widgets με την διαδικασία drag and drop και ένωσης μεταξύ τους, διαδικασία που είναι περισσότερο φιλική προς τον χρήστη. Συνδυάζοντας λοιπόν τον κατάλογο των widgets στην ιστοσελίδα του προγράμματος και με την drag and drop χρήση, δημιουργεί μια θετική πρώτη εντύπωση.

Το **Rapidminer** που είναι σήμερα και το πιο διαδεδομένο είναι περισσότερο φιλικό προς τον χρήστη από τα προηγούμενα. Ανοίγοντας την εφαρμογή, κάτω

αριστερά έχει μια ενότητα “How to start” που οδηγεί στην ιστοσελίδα http://rapid-i.com/content/view/189/212/, όπου υπάρχουν video tutorials που οδηγούν τον χρήστη βήμα-βήμα στο πως μπορεί να χειριστεί το εργαλείο. Αυτό από μόνο του είναι πολύ χρήσιμο αν και το πρόγραμμα παρέχει στον χρήστη και tutorial βήμα-βήμα μέσα στο πρόγραμμα. Επιλέγοντας από το μενού Help την επιλογή Rapidminer tutorial ανοίγει ένα νέο παράθυρο που καθοδηγεί τον νέο χρήστη και του παρουσιάζει τις δυνατότητες του εργαλείου. Στο μενού Help υπάρχουν επίσης και τα video tutorials και το link για το forum του εργαλείου όπου μπορεί ο χρήστης να συμμετέχει σε συζητήσεις πάνω στο αντικείμενο ή να θέσει ερωτήσεις για το εργαλείο όπου υπάρχει άμεση ανταπόκριση. Τέλος, το Rapidminer μπορεί να αναβαθμιστεί οποτεδήποτε σε κάποια νέα έκδοση από το μενού Help και επιλέγοντας Update Rapidminer.

6.2 Τεχνικά προβλήματα

Μέσα από την χρήση των εργαλείων παρατηρήθηκαν κάποια προβλήματα τεχνικής φύσεως που θα πρέπει να αναφερθούν. Στο **Weka,** το πρόβλημα που αντιμετωπίσαμε κατά την χρήση του ήταν πολύ σημαντικό και αν δεν επιλυόταν δεν θα επιτρεπόταν η χρήση του για μεγάλα αρχεία δεδομένων. Επειδή το πρόγραμμα τρέχει εξολοκλήρου στην μνήμη Ram, δεσμεύει κατά την εκκίνηση ένα κομμάτι της για να τρέξει. Η default επιλογή του προγράμματος όμως ήταν αρκετά μικρή (της τάξεως των 64mb) με αποτέλεσμα όταν τρέχαμε την συσταδοποίηση να μας βγάζει το Out of memory μήνυμα της εικόνας 5.5 (βλ σελ 53). Το πρόγραμμα στην συνέχεια έκλεινε με αποτέλεσμα ό,τι δουλειά ή παραμετροποίηση είχε κάνει ο χρήστης να χαθεί. Στο troubleshooting του επίσημου site του Weka προτεινόταν να αυξηθεί στην Java η δέσμευση μνήμης έτσι ώστε να μπορεί να χειριστεί και τα αυξημένα δεδομένα. Μέσω όμως ενός forum στο διαδίκτυο βρέθηκε η λύση. Έπρεπε να αλλαχτεί στο αρχείο RunWeka.ini που είναι το αρχείο των settings του προγράμματος το maxheap που ορίζει την αρχική δέσμευση RAM. Ορίζοντάς το στα 1024mb το πρόβλημα λύθηκε και δεν επανεμφανίστηκε το error message. Θα έπρεπε όμως να δεσμεύεται ως προεπιλογή περισσότερη RAM για να μην αντιμετωπίζει ο χρήστης τέτοιο πρόβλημα.

Στο **Orange** παρατηρήθηκαν κάποια προβλήματα (σπάνια) όπου κατά την εκτέλεση κάποιου αλγορίθμου ή κάνοντας κάποιες επιλογές, το πρόγραμμα

σταματούσε να αποκρίνεται με αποτέλεσμα το «βίαιο κλείσιμο» του προγράμματος και χάνονταν έτσι όλη η προηγούμενη δουλειά. Επίσης στα διάφορα γραφήματα αλλάζοντας κάποιες επιλογές, άλλαζαν τα χρώματα στο σχήμα με αποτέλεσμα να μην ανταποκρίνεται πχ η σωστή συστάδα με το χρώμα της. Το πρόβλημα διορθωνόταν αν ξαναρύθμιζες το γράφημα απ την αρχή

Τέλος με το **Rapidminer** δεν παρουσιάστηκαν κάποια τεχνικά προβλήματα.

6.3 Δεδομένα και Συσταδοποίηση

Σε αυτή την ενότητα θα δούμε ποιο εργαλείο εισάγει με τον πιο απλό τρόπο και εμφανίζει τα αρχικά δεδομένα με τον πληρέστερο τρόπο.

Το **Weka** με μερικές πολύ απλές κινήσεις επιτρέπει την εισαγωγή των δεδομένων μέσω της επιλογής *open file…* στην καρτέλα Preprocess στον περιβάλλον του Explorer. Επιλέγοντας την κάθε μια από τις μεταβλητές, δεξιά εμφανίζονται τα στατιστικά στοιχεία και η κατανομή σε ένα ασπρόμαυρο γράφημα. Επίσης, επιλέγοντας τις μεταβλητές που δεν θέλουμε και πατώντας το κουμπί *remove* αφαιρούνται οι μεταβλητές από τα δεδομένα πολύ εύκολα και απλά. Ένα αρνητικό που παρατηρείται είναι ότι όταν επιλέγουμε φίλτρο για τα δεδομένα, δεν υπάρχει επεξήγηση για το τι κάνει το κάθε φίλτρο με αποτέλεσμα να πρέπει να ψάξεις στην ιστοσελίδα του προγράμματος για περισσότερες πληροφορίες. Σε ότι έχει να κάνει με την συσταδοποίηση, το εργαλείο έχει 12 διαφορετικούς αλγορίθμους. Και πάλι όμως πρέπει, αν δεν ξέρει κάποιος, να ψάξει στην ιστοσελίδα του εργαλείου για να πάρει επεξηγήσεις για το τι κάνει ο κάθε αλγόριθμος. Οι 12 αυτοί αλγόριθμοι είναι: CLOPE, Cobweb, DBScan, EM, FarthestFirst,FilteredClusterer, HierarchicalClusterer,MakeDensityBasedClusterer, OPTICS, sIB, SimpleKMeans, Xmeans.

Στο **Orange** η εισαγωγή των δεδομένων γίνεται λίγο πιο πολύπλοκα αλλά χωρίς ιδιαίτερη δυσκολία και πάλι. Απαιτείται η εισαγωγή με την σειρά του *Open File* widget και μετά του *Select Attributes* widget. Στο πρώτο εισάγουμε το αρχείο και στο δεύτερο επιλέγουμε τις μεταβλητές που θέλουμε. Για να δούμε τα στατιστικά όπως στο Weka πρέπει να βάλουμε το *Distributions* για τις κατανομές συχνοτήτων και το *Attribute Statistics* για τα στατιστικά. Το μεγάλο αρνητικό για το Orange είναι ότι δεν υπάρχει ποικιλία αλγορίθμων για συσταδοποίηση. Περιέχει

μόνο 2 αλγορίθμους (Hierarchical Clustering, k-Means Clustering) για συσταδοποίηση κάνοντας έτσι το εργαλείο πολύ φτωχό σε αυτό τον τομέα.

Στο **Rapidminer** για την εισαγωγή των δεδομένων είναι πιο πολύπλοκο αλλά και πιο αναλυτικό. Εισάγουμε το Read CSV εικονίδιο και επιλέγουμε το *Import Configuration Wizard...* και ανοίγει ένας οδηγός που καθοδηγεί τον χρήστη βήμα – βήμα από την επιλογή του αρχείου, στον ορισμό του χαρακτήρα που χωρίζει τις στήλες στο αρχείο, τον ορισμό του τύπου των μεταβλητών έχοντας μια προεπισκόπηση των πρώτων 100 εγγραφών του αρχείου. Υπάρχει ακόμα και αυτόματη επιλογή των τύπων από το κουμπί *Guess value types* που ορίζει αυτόματα τους τύπους των μεταβλητών. Τέλος, μπορεί να απο-επιλεγεί οποιαδήποτε από τις μεταβλητές έτσι ώστε να μην την λάβει υπόψη του. Ως προς την ποικιλία αλγορίθμων στην συσταδοποίηση, έχει 12 αλγορίθμους δικούς του (k-means, k-means (Kernel), k-means (fast), k-Medoids, DBSCAN, Expectation Maximization Clustering, Support Vector Clustering, Random Clustering, Agglomerative Clustering, Top Down Clustering, Flatten Clustering, Extract Cluster Prototypes) συν άλλους 7 του Weka, (W-SimpleKMeans, W-CLOPE, W-Cobweb, W-EM, W-FarthestFirst, W-XMeans, W-sIB) και μάλιστα πηγαίνοντας το ποντίκι πάνω από κάθε έναν αλγόριθμο ή επιλέγοντάς τον υπάρχει αναλυτική επεξήγηση του τι κάνει ο αλγόριθμος.

Στον πίνακα 6.1 βλέπουμε τις διαφορές ως προς την ποικιλία αλγορίθμων συσταδοποίησης από πρόγραμμα σε πρόγραμμα.

Πίνακας 6.1: Αλγόριθμοι συσταδοποίησης ανά πρόγραμμα.

	Weka	Orange	Rapidminer
Αλγόριθμοι συσταδοποίησης	12	2	12+7

6.4 K-means και Παραμετροποίηση

Σε αυτή την ενότητα θα συγκριθούν οι παράμετροι που υπάρχουν για τον k-means ανά πρόγραμμα.

Στο **Weka** υπάρχουν 7 διαφορετικές παράμετροι που μπορεί ο χρήστης να καθορίσει την τιμή τους έτσι ώστε να εφαρμόσει τον αλγόριθμο όπως επιθυμεί. Οι

παράμετροι αυτές αναφέρονται στον υπολογισμό των τυπικών αποκλίσεων, στην επιλογή του μέτρου της απόστασης, στην διαχείριση των missing values, στον μέγιστο αριθμό επαναλήψεων του αλγόριθμου, στο πλήθος των συστάδων, στην επιλογή της σειράς των εγγραφών και τέλος στον τυχαίο αριθμό που θα καθορίσει τα αρχικά κέντρα.

Το **Orange** είναι σαφέστατα φτωχότερο από το Weka καθώς έχει 4 μόνο παραμέτρους: το πλήθος των συστάδων, την επιλογή τρόπου του μέτρου απόστασης, τον τρόπο εκκίνησης των αρχικών κέντρων και το πλήθος των επαναλήψεων.

Το **Rapidminer** είναι κι αυτό φτωχότερο στην παραμετροποίηση από το Weka. Έχει 4 παραμέτρους για τον αλγόριθμο του προγράμματος: το πλήθος των συστάδων, τον μέγιστο αριθμός εκτελέσεων, τον μέγιστο αριθμός επαναλήψεων και το αν θέλει ο χρήστης, για την επιλογή των αρχικών κέντρων, να χρησιμοποιήσει έναν τοπικό τυχαίο αριθμό seed και αν ναι να τον ορίσει ο χρήστης. Για τον αλγόριθμο του Weka που έχει το Rapidminer ισχύει ότι και στο Weka, έχει δηλαδή την ίδια ακριβώς παραμετροποίηση με το Weka.

Πίνακας 6.2: Αριθμός παραμέτρων για τον k-means ανά πρόγραμμα

	Weka	Orange	Rapidminer
Αριθμός παραμέτρων	7	4	4

6.5 Αποτελέσματα

Σε αυτή την ενότητα θα συγκριθούν τα αποτελέσματα που επέστρεψε το κάθε πρόγραμμα.

Τα αποτελέσματα ανάμεσα στο **Weka** και το **Orange** είναι παρεμφερή. Οι τρεις συστάδες και στα δυο είναι περίπου ισομεγέθεις και τα προφίλ του πελάτη σε κάθε συστάδα είναι περίπου τα ίδια. Έχουν χωριστεί ανά οικονομική κατάσταση, δηλαδή μισθό, καταθέσεις σε τρεις διαφορετικές ας το πούμε «τάξεις». Υπάρχουν οι χαμηλόμισθοι (Weka-380, Orange-250), τα μεσαία εισοδήματα (Weka-1968, Orange-1653) και οι υψηλόμισθοι (Weka-3953, Orange-3813)(Τα νούμερα στις παρενθέσεις αποτελούνε μέσες τιμές). Το **Rapidminer** απ' την άλλη, λόγω της ιδιαιτερότητας που έχει ο δικός του αλγόριθμος να μην δέχεται ποιοτικές

μεταβλητές και την υποχρεωτική μετατροπή των ποιοτικών μεταβλητών σε αριθμητικές, είχε όπως είναι φυσικό διαφορετικά αποτελέσματα. Οι τρεις συστάδες δεν είναι ισομεγέθεις καθώς η πρώτη έχει 14188, η δεύτερη 1780 και η τρίτη 27. Τα εισοδήματα έχουν μεν μια διαβάθμιση από συστάδα σε συστάδα αλλά δεν ήταν ο κύριος παράγοντας για τον διαχωρισμό τους.

Πίνακας 6.3: Τα αποτελέσματα του αλγορίθμου k-means στον Rapidminer

	Συστάδα 0	Συστάδα 1	Συστάδα 2
Μισθός	1567	5744	7915
Κρατικά ομόλογα	3760	13669	444487

Όπως βλέπουμε και στον πίνακα 6.3 τα κρατικά ομόλογα είναι πιο διακριτές ποσότητες απ' ότι ο μισθός. Σε κάθε περίπτωση, πάντως, το ότι δεν μας επιτρέπεται να χρησιμοποιήσουμε αυτούσιες ποιοτικές μεταβλητές, αλλοίωσε τα αποτελέσματά μας και είναι μεγάλος αρνητικός παράγοντας. Τέλος τα αποτελέσματα του Weka k-means στο Rapidminer δεν παρατίθενται με συγκεντρωτικά στατιστικά στοιχεία έτσι ώστε να μπορέσουμε να βγάλουμε συμπεράσματα. Το μόνο που μπορούμε να πούμε είναι ότι οι τρεις συστάδες είναι περίπου ισομεγέθεις.

6.6 Εμφάνιση αποτελεσμάτων

Σε αυτή την ενότητα θα συγκριθούν οι τρόποι που παρέχονται στον χρήστη από το κάθε πρόγραμμα για να εμφανιστεί ή να επεξεργαστεί τα αποτελέσματα πχ με γραφήματα.

Στο **Weka** υπάρχει ένας Plot matrix που μπορούμε να θέσουμε στους άξονες χ και ψ δυο μεταβλητές και βάζοντας με χρώμα μια τρίτη (πχ την συστάδα) να βγάλουμε συγκεκριμένα συμπεράσματα ως προς την κατανομή των στοιχείων του αρχείου της μελέτης περίπτωσης. Αυτό είναι και το μόνο γράφημα που παρέχεται από το Weka.

Από την άλλη στο **Orange** εκτός από το Scatter plot το οποίο έχει την ίδια λειτουργία με το Plot matrix του Weka, έχει και το Distributions που μπορούμε να δούμε τις κατανομές ανά συστάδα και επίσης την συχνότητα και τις πιθανότητες

εμφάνισης ανά συστάδα κάποιων ποσοτήτων, πχ πιθανότητα εμφάνισης υψηλού μισθού στην πρώτη συστάδα.

Ο **Rapidminer** στο θέμα της εμφάνισης των αποτελεσμάτων με διάφορα γραφήματα είναι με διαφορά το πιο πλήρες πρόγραμμα από τα 3. Στα αποτελέσματα του kmeans έχει έναν Plotter ο οποίος μπορεί να αλλάξει σε 32 διαφορετικά γραφήματα. Υπάρχει το κλασικό scatter plot που έχουν και τα άλλα δυο εργαλεία, το γράφημα πίτας, γράφημα σειρών, γράφημα πυκνότητας, ιστογράμματα και τα περισσότερα απ αυτά μπορούν εκτός από δυο διαστάσεις να αναπαρασταθούν και σε τρισδιάστατη μορφή.

ΑΝΑΦΟΡΕΣ

[1] http://users.sch.gr/apouliassis/glosses2/dedomena.htm

[2] http://el.wikipedia.org/wiki/Πληροφορία

[3] http://en.wikipedia.org/wiki/Data_mining

[4] W. Frawley and G. Piatetsky-Shapiro and C. Matheus (Fall 1992). "Knowledge Discovery in Databases: An Overview". AI Magazine: pp. 213-228. ISSN 0738-4602. Ανάκτηση από http://www.datamining.gr/el/whatisdatamining.html

[5] D. Hand, H. Mannila, P. Smyth (2001). Principles of Data Mining. MIT Press, Cambridge, MA. ISBN 0-262-08290-X. Ανάκτηση από http://www.datamining.gr/el/whatisdatamining.html

[6] Ellen Monk, Bret Wagner (2006). Concepts in Enterprise Resource Planning, Second Edition. Thomson Course Technology, Boston, MA. ISBN 0-619-21663-8. OCLC 224465825. Ανάκτηση από http://www.datamining.gr/el/whatisdatamining.html

[7] http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm

[8] http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html

[9] http://en.wikipedia.org/wiki/RapidMiner

[10] Ian H. Witten; Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham (1999). "Weka: Practical Machine Learning Tools and Techniques with Java Implementations". *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*. pp. 192–196. Ανάκτηση από http://en.wikipedia.org/wiki/Weka_(machine_learning)

[11] Gregory Piatetsky-Shapiro (2005-06-28). "KDnuggets news on SIGKDD Service Award 2005". Ανάκτηση από http://en.wikipedia.org/wiki/Weka_(machine_learning)

[12] Overview of SIGKDD Service Award winners". 2005. Retrieved 2007-06-25. Ανάκτηση από http://en.wikipedia.org/wiki/Weka_(machine_learning)

[13] http://en.wikipedia.org/wiki/Weka_(machine_learning)

[14] http://en.wikipedia.org/wiki/Orange_(software)

[15] http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf

[16] Pang-Ning Tan, Michael Steinbach,Vipin Kumar, “Introduction to Data Mining”, Addison Wesley,2006, Ανάκτηση από http://www-users.cs.umn.edu/~kumar/dmbook/index.php

[17] http://www.differencebetween.com/difference-between-hierarchical-and-vs-partitional-clustering/

[18] http://en.wikipedia.org/wiki/Cluster_analysis

[19] http://www.slideshare.net/pierluca.lanzi/lecture-01-data-mining

[20] Karl Rexer, Heather Allen, & Paul Gearan (2010) *2010 Data Miner Survey Summary*, presented at Predictive Analytics World, Oct. 2010. Ανάκτηση από http://en.wikipedia.org/wiki/Data_mining

[21] http://www.redbooks.ibm.com/redbooks/pdfs/sg246879.pdf